Flexible Corticospinal Control of Muscles

Najja J. Marshall

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

Abstract

Flexible Corticospinal Control of Muscles

Najja J. Marshall

The exceptional abilities of top-tier athletes – from Simone Biles' dizzying gymnastics to LeBron James' gravity-defying bounds – can easily lead one to forget to marvel at the exceptional breadth of everyday movements. Whether holding a cup of coffee, reaching out to grab a falling object, or cycling at a quick clip, every motor action requires activating multiple muscles with the appropriate intensity and timing to move each limb or counteract the weight of an object. These actions are planned and executed by the motor cortex, which transmits its intentions to motoneurons in the spinal cord, which ultimately drive muscle contractions. A central problem in neuroscience is precisely how neural activity in cortex and the spinal cord gives rise to this diverse range of behaviors. At the level of spinal cord, this problem is considered to be well understood. A foundational tenet in motor control asserts that motoneurons are controlled by a single input to which they respond in a reliable and predictable manner to drive muscle activity, akin to the way that depressing a gas pedal by the same degree accelerates a car to a predictable speed. Theories of how motor cortex flexibly generates different behaviors are less firmly developed, but the available evidence indicates that cortical neurons are coordinated in a similarly simplistic, well-preserved manner. Yet a potential complication for both these old and new theories are the relative paucity of diverse behaviors during which motor cortex and spinal motoneurons have been studied. In this dissertation, I present results from studying these two neuronal populations during a broader range of behaviors than previously considered. These results indicate, in essence, that diverse behaviors involve greater complexity and flexibility in cortical and spinal neural activity than indicated by current theories.

Table of Contents

List of Figures iv

Acknowledgments vi

Dedication x

Chapter 1: Introduction 1

1.1 Outline of dissertation 2

1.2 Models of motor cortex 3

1.3 Low-dimensional cortical manifolds in motor control 5

1.4 Motor units and their inputs 9

1.5 Motor unit diversity 12

1.6 Orderly motor unit recruitment 14

1.7 Apparent violations of orderly MU recruitment 18

Chapter 2: Flexible Neural Control of Motor Units 23

2.1 Introduction 24

2.2 Results 25

2.2.1 Pac-Man Task and EMG recordings 25

2.2.2 Motor unit activity during behavior 28

2.2.3 Cortical perturbations . . . 30
2.2.4 State space predictions of rigid control . . . 31
2.2.5 Latent factor model . . . 33
2.2.6 Neural degrees of freedom . . . 36
2.3 Discussion . . . 39
2.4 Methods . . . 41
2.4.1 Data acquisition . . . 41
2.4.2 Data processing . . . 47
2.4.3 Data Analysis . . . 49
2.5 Supplementary Figures . . . 57
2.6 Supplementary Materials . . . 62
2.6.1 EMG Signal Decomposition . . . 62
2.6.2 Optimal Motor Unit Recruitment . . . 73

Chapter 3: High-dimensional neural manifolds for complex muscle control . . . 83
3.1 Introduction . . . 84
3.2 Results . . . 86
3.2.1 Task and behavior . . . 86
3.2.2 Single-neuron response features . . . 87
3.2.3 Neural correlates with behavior . . . 89
3.2.4 Population structure . . . 91
3.2.5 Neural subspace alignment . . . 94
3.3 Discussion . . . 96

3.4 Methods 99
3.4.1 Data acquisition 99
3.4.2 Data processing 101
3.4.3 Data Analysis 102
3.5 Supplementary Figures 107

Chapter 4: Conclusion 111
4.1 Summary 111
4.2 Future Directions 113
4.2.1 Optimal motor unit recruitment 113
4.2.2 Sources of high dimensionality in cortex 115
4.2.3 Task elaborations 115

References 117

List of Figures

1.1 Projections of the neural population response 5

1.2 Neural Data Manifolds 8

1.3 Muscle proprioceptive feedback and spinal networks 11

1.4 Orderly motor unit (MU) recruitment 16

2.1 Experimental setup and MU spikes 26

2.2 Example MU responses 29

2.3 State space predictions for rigid and flexible MU control 32

2.4 Latent Factor Model 35

2.5 Quantifying neural degrees of freedom 37

S2.6 Force profiles 57

S2.7 Example MU responses and waveforms across muscle lengths 58

S2.8 MU displacement 59

S2.9 Motor neuron pool (MNP) dispersion 60

S2.10 Example primary motor cortex (M1) neuron responses 61

S2.11 Twitch responses 79

S2.12 Optimization predictions 80

S2.13 Recruiting fast MUs increases force accuracy 81

3.1 Task and neural recordings 86

3.2 Example primary motor cortex (M1) neuron responses 88

3.3 Encoder models of single-neuron responses 90

3.4 Bias-variance tradeoff in readout dimensions 91
3.5 M1 activity visualized in a low-dimensional subspace 92
3.6 M1 activity in three different subspaces 93
3.7 Subspace alignment across conditions 95
3.8 Variance explained across conditions 96
S3.9 Multi-muscle EMG activity 107
S3.10 Bias-variance tradeoff in motor readouts 108
S3.11 Motor unit variance explained across conditions 109
S3.12 Trajectory tangling 110

Acknowledgements

What a ride. This has undoubtedly been the most challenging experience in my life, but also one of the most rewarding. As I reflect upon how I made it to this point, I have to acknowledge all of the encouraging, nurturing, and generally wonderful people that made it all possible.

Thank you, mom, dad, and A.D. for your love and support. Mom and dad, thank you for investing so heavily in my education and for always taking an interest in my scientific endeavors. Thanks for the weekly phone calls ever since I left for college and for all the care packages. Staying connected with you all has certainly helped me stay grounded.

Thank you, to all of the teachers and mentors that I have been so incredibly fortunate to have had. Thank you, Judy Tisdale for instilling a passion to understand the world around us; Sandra Mueller for showing me that science can be fun; and Jerry Taylor for encouraging me to pursue an interest in physics. Thank you, Brian Connor and Dan Barton for the first opportunity to embark on a *long*-term project and for sticking with me all the way through. I don't think any of us expected to spend four years building a pool table, but I firmly believe that the skills I gained proved invaluable for building the apparatuses for my scientific experiments years later. Thank you, Daniel ('Coach') Harris for being a role model and someone I could always turn to for guidance. Thanks for teaching me how to lead and the importance of teamwork. And thanks for the constant reminder: "whether you think you can, or think you can't – you're right".

Thank you, Prof. John Beggs for welcoming me into your laboratory. Thanks for showing me what it means to conduct scientific research. And thank you, Dr. Nick Timme for taking me under your wing, for being a fantastic collaborator, and for teaching me MATLAB.

Thank you, Profs. Babak Seradjeh and Kent Orr for challenging me to push myself and for your continued guidance. Thanks for teaching me to think critically and deeply.

Thank you, to my PhD cohort: Sebi Rolotti, Ella Batty, Macayla Donegan, Laura Long, Rebecca Vaadia, Rachel Clary, Cat Braine, Neeli Mishra, Ali Kaufman, Melina Tsitsiklis, Rick Warren, Melissa Lee, Dan Kato, and Rikki Rabinovich. You are all wonderful people. I am so glad to have embarked on this journey with you, and that we were able to gather for cocktail parties and ski trips along the way.

Thank you, to my PhD advisors, Profs. Mark Churchland and Larry Abbott. Thank you, Larry for your keen insights and wisdom. Your ability to quickly assess the crux of a problem in seemingly any domain is uncanny, and your feedback has always proven invaluable. Thank you, Mark for being an exceptionally dedicated advisor and mentor. Thanks for not just guiding my experiments, but for donning your lab coat and working with me, side by side. Thanks for being so generous with your time – I cannot possibly count the number of times a 'simple' question turned into an hours-long discussion. Thanks for your extreme attention to detail, for your willingness to work through every piece of written text (be it a manuscript or poster abstract) line by line. Thanks for fostering such a pleasant and inclusive environment to do science. And thanks for demonstrating that being a fantastic scientific advisor extends beyond the science itself. Thanks for the lab retreats and holiday parties, for the discussions about politics or weightlifting, and thanks for all the cocktail recipes. I could not have asked for a better advisor.

Thank you, to the members of the Churchland Lab, past and present, for being amazing colleagues and even better friends. Thank you, Yana Pavlova for all of your help with the 'little guys' and for providing us with a constant supply of coffee and chocolate. Thanks for ensuring that we always take the time to get together for happy hours, and for never missing a birthday. Thank you,

Cora Ames for building 'Jumanji' with me and for being a great scientific role model. Thank you, Andrew Zimnik for your wealth of scientific insight. Thanks for being my conference partner-in-crime, lifting buddy and fellow political junkie. Thank you to everyone else who has helped make the lab a place I eagerly looked forward to coming to every morning: Sean Perkins, Abby Russo, Karen Schroeder, Antonio Lara, Jeff Seely, Eric Trautmann, Elom Amematsro, Saurabh Vyas, and Elijah Aliyari.

Thank you, to the members of the Abbott Group and all of the incredible people in the Theory Center. Thank you, Jacob Portes and Matthias ('Gucky') Christenson for convincing me to learn python – I only wish you had done so sooner. Thank you, Salomon Muller, Ramin Khajeh, Denis Turcu, Dan Tyulmankov, Marjorie Xie, Ching Fang, and David Clark for all of your feedback on my presentations. Thank you, Prof. John Cunningham, for your support and helpful advice. Thank you, Sean Bittner, Chris Cueva, Rainer Engelken, Laureline Logiaco, Taiga Abe, and Dan Biderman. And thank you to the former graduate students and post docs who helped me immensely, particularly in the early days: Drs. Patrick Kaifosh, Brian DePasquale, and Marcus Benna.

Thank you, to my collaborators. Thank you, Sean Perkins for showing me the ropes around the lab. Thanks for helping me design my first experiments and for teaching me Simulink, and thanks for being my point person for math and programming questions. Thank you, Josh Glaser for sticking with us for so long through such a demanding project, and thank you for providing so much valuable feedback and insight along the way. Thank you, Eric Trautmann for ushering us into the next generation of neural recording technologies, and thanks for helping spike sort my data sets. Thank you, Elom Amematsro for your uncannily adept analyses and modeling work.

Thank you, to my thesis committee: Profs. Rui Costa, Daniel Wolpert, Liam Paninski, and Andrew Pruszynski. Thanks for your insightful questions and helpful feedback. Thank you, Rui for being a guiding light for Zuckerman. Thanks for encouraging everyone to take a break and have

a bit of fun. And thanks for including my voice in the effort to make the Institute a more diverse, equitable and inclusive place.

Thank you, to my NYC family, and to all of my friends, old and new. Thank you to my cousins, Aisha Russell and Marjani Jones, for always coming through with the activities and for keeping this place interesting. Thank you, Adam Kallaus for being my 'day one'. Thanks for staying such a great friend, and for always proudly introducing me as your neuroscientist friend. Thank you, Chris Hanson for slogging through physics problem sets together and for turning me into a LaTeX-olyte. And of course, thanks for encouraging me to apply to Columbia – I very likely would not be here if not for you. Thank you, Andrew McLaren for embarking on the journey from IUB to NYC together, and for being such a great roommate for so many years. Thank you, to the 'squad': Peter Mercado, Tony Huggins, Ian Reyes, Jacobi Holland, Maurice Blackmon, Cornel Duhaney, and Ron Norman. You are all an amazing group of guys. Thanks for being people I can talk to about science or finance, and also count on for a weekend of shenanigans. Thank you to all of my other wonderful NYC friends: Joy and June Harewood, Barry Hoy, Teraj Allen, Alyssa Curan, and Anthony Diaz-Santana.

Thank you, to my partner and best friend, Cicely Shillingford. Thanks for being an equally passionate scientist, foodie, fitness enthusiast, and party host. Thanks for the living room practice talks and writing advice. I am in awe of your work ethic and your commitment to balance. Thanks for pulling me out of my work hole and ensuring that we take breaks or get away for a weekend. Thank you for your kindness and generosity, for your unbridled love and support.

For my parents, Pam & Vincent

Chapter 1: Introduction

The primary motor cortex (M1) contains 1.3 billion neurons[1]. Principally responsible for planning and executing voluntary movement[2], M1 drives motor functions though the excitation and inhibition of α-motoneurons in the spinal cord. Each motoneuron innervates and controls a set of muscle fibers; the neuron and its fibers are collectively called a motor unit (MU)[3]. MUs constitute the functional atoms of the neuromuscular system, with hundreds of MUs controlling the activation of each muscle[4]. Fundamentally, this dissertation concerns the control and coordination of these two neuronal populations – M1, and MUs dedicated to a particular muscle – for flexibly driving behavior.

The Soviet neurophysiologist Nikolai Bernstein defined coordination as, "a problem of mastering the very many degrees of freedom involved in a particular movement – of reducing the number of independent variables to be controlled[5]." This concept, of reducing degrees of freedom to simplify control, underlies emerging theories of M1 and the canonical understanding of MUs. It would certainly be outlandish to suggest that every individual neuron or MU operates independently of its companions; the issue at hand, rather, is a matter of degree. How many neural degrees of freedom underlie M1 or MU activity? If not a billion, then millions? If not hundreds, perhaps dozens? A recent theory, based on multiple lines of evidence, posits that M1 flexibly recombines a few (order tens) well-preserved neural modes to generate a diverse range of behaviors[6]. Similarly, nearly a century of research has established the perception that MUs are driven by one degree of freedom as a principle of neural science[7]. These perspectives argue for a simplistic and rigid view of motor control.

1.1 Outline of dissertation

In this dissertation, I present work revealing greater complexity and flexibility in M1 and MU control than previously considered. In the remainder of **chapter 1**, I provide additional context for the present understanding of these neuronal populations. I review models of how M1 generates movement in **section 1.2** and the evidence suggesting that M1 may rely on relatively few degrees of freedom to produce different behaviors in **section 1.3**. I then review MUs and their inputs in **section 1.4**, the various forms of MU diversity in **section 1.5**, the evidence underlying the canonical description of MU control as rigid and one-dimensional in **section 1.6**, and the literature on apparent exceptions to the rigid description of MU control in **section 1.7**.

In **chapter 2**, I present my work on flexibility in MU control. I introduce a novel behavioral paradigm that facilitates investigations of MU activity across a diverse range of behaviors. Using this paradigm, I present results indicating that multiple degrees of freedom underlie MU control. Motivated by our empirical findings, I describe a simple model that supports an alternative hypothesis for MU control separate from the canonical description.

In **chapter 3**, I present my work on complexity in M1. Using the same behavioral paradigm and high-density recordings of neural activity, I present results revealing that M1 produces different behaviors by employing many neural degrees of freedom, at least an order of magnitude higher than previous estimates.

In **chapter 4**, I provide concluding remarks and discuss potential future directions.

1.2 Models of motor cortex

Conceptual frameworks for how motor cortex produces movement have evolved considerably over several decades. Initial hypotheses posited that motor cortex encodes high-level movement parameters in the patterns of single-cell discharge. Under this *representational* hypothesis, the firing rate of neuron i at time t is given by

$$r_i(t) = f_i(\Theta(t)) \tag{1.1}$$

where f_i is the 'tuning' function for neuron i, and Θ is a set of general parameters. Across various behavioral tasks, including flexion-extension rotations of the wrist[8–14], ballistic point-to-point reaches[15–19], or continuous arm movements[20], single-neuron firing rates correlate well with endpoint force or its derivative[8–11], muscle activity[12,13], joint position[12], movement direction[12,14–17], or linear combinations of hand kinematics[18–20]. These parameters were all proposed as potential candidates for Θ (**eq.** (1.1)). Yet it is possible to find individual neurons that correlate well with any conceivable movement parameter[12], and representational models do not explain how neural correlates with behavior causally drive movement. Further complicating matters, single-cell tuning functions (f_i) vary with arm posture[21], movement speed and time[22]. Additionally, single-cell correlations with kinematic parameters incidentally emerge from a mechanistic model in which cortical neurons drive arm movements by controlling muscle activity[23]. These confounding observations for the representational hypothesis argued for an alternative model of motor cortex.

An alternative view of motor cortex proposed that it operates as a dynamical system to generate the appropriate patterns of muscle activity that ultimately drive movement[24,25]. Under this *dynamical systems* hypothesis, the firing rate of a population of N neurons, $\mathbf{r}(t) = [r_1(t), r_2(t), \ldots, r_N(t)]$, evolves according to some set of lawful dynamics:

$$\dot{\mathbf{r}}(t) = f(\mathbf{r}(t)) + \mathbf{u}(t) \tag{1.2}$$

where f is some function and $\mathbf{u}(t)$ an external input. As a 'population-level' view, by definition, the dynamical systems hypothesis was largely facilitated by methods for visualizing the activity of many neurons in a low-dimensional space, such as principal component analysis (PCA). PCA identifies an orthogonal set of dimensions that capture the largest (highest variance) signals in some data. If $R \in \mathbb{R}^{(\sum_c T_c) \times N}$ contains the response of each neuron across multiple experimental conditions (e.g., reach directions), where T_c is the duration of condition c, then principal components (PCs) of R can be obtained through an eigendecomposition of its $N \times N$ covariance matrix, Σ:

$$U, \Lambda, U^\top = \Sigma \tag{1.3}$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ and $U = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_N \end{bmatrix}$. Each $\mathbf{u}_i$ is an eigenvector (PC), sorted in descending order by its corresponding eigenvalue (λ_i). The leading PCs of R thus capture the largest patterns of neural coactivation. A common technique for visualizing those patterns involves applying PCA (or related variations) to neural activity, projecting activity onto its leading components (e.g., $[R\mathbf{u}_1, R\mathbf{u}_2]$), then plotting the projections against one another.

Analyzing and visualizing the largest neural signals during behavior has proven invaluable for assessing competing models of motor cortex. One of the most salient features that emerges from low-dimensional projections of motor cortex activity is strong and robust rotational dynamics in the largest neural signals during reaching[24,26–30] (**fig. 1.1**) and cycling[31]. Rotations do not arise trivially from smoothed neural responses or correlations across time, conditions, or neurons[32]. Recurrent neural networks (RNNs) develop rotational dynamics when trained to produce reach-related muscle activity[33]. Rotations also appear in reach-velocity-tuned models of neural activity, provided that they include sufficient variability in neuron-kinematic latencies[34]. However, rotations in velocity-tuned models persist after permuting the responses of each neuron across conditions while preserving the overall neural covariance structure, whereas the same control destroys rotations in RNNs and motor cortex data, indicating that rotational dynamics in motor cortex depend on the

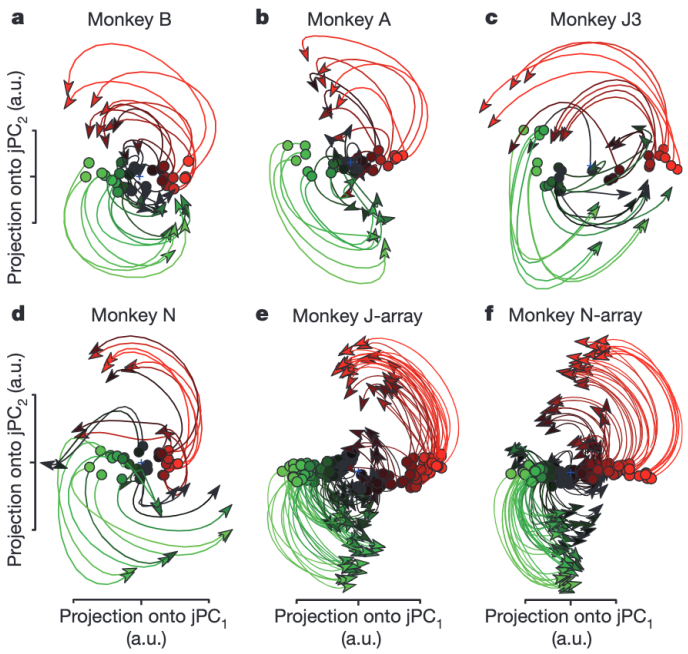


Figure 1.1: Projections of the neural population response. **a**, Projection for monkey B (74 neurons; 28 straight-reach conditions). Each trace (one condition) plots the first 200 ms of movement-related activity away from the preparatory state (circles). Traces are coloured on the basis of the preparatory state projection onto jPC1. a.u., arbitrary units. **b**, Projection for monkey A (64 neurons; 28 straight-reach conditions). **c**, Monkey J, data set 3 (55 neurons; 27 straight- and curved-reach conditions). **d**, Monkey N (118 neurons; 27 straight and curved-reach conditions). **e**, Monkey J-array (146 isolations; 108 straight and curved-reach conditions). **f**, Monkey N-array (218 isolations; 108 straight and curved-reach conditions). (Reproduced from Churchland *et al.*, 2012, with permission).

underlying structure of neural activity across conditions[34]. Large rotational signals increase the noise robustness of RNNs, and muscle activity can be linearly read out from smaller signals riding on top of rotational dimensions[31]. These findings indicate that motor cortex is better explained as a dynamical system that generates movement, rather than as a collection of neurons that represent high-level movement parameters.

1.3 Low-dimensional cortical manifolds in motor control

The broad motor repertoire of humans and other creatures along with the understanding of motor cortex as a movement-generating dynamical system raises an important question: how does motor cortex flexibly generate different behaviors? On the one hand, motor cortex could operate

as a "generalist", simply reusing a small number of neural modes (patterns of neural coactivation) to produce different movements; on the other hand, motor cortex might operate as a "specialist", relying on entirely unrelated neural modes to subserve different behaviors. This question really has two components: how many cortical modes are involved in motor control and how well preserved are those modes across different behaviors? The first component has received considerably more attention than the second, though the available evidence for both tends to favor the generalist view of motor cortex.

A common method for estimating the number of neural modes that underlie behavior involves simply counting the number of PCs that capture the most variance in neural activity. This can be achieved by identifying where the proportional cumulative variance explained by each neural PC exceeds some threshold. That is, given Λ (**eq.** (1.3)) and some threshold α, the putative number of neural modes is the n for which

$$\frac{\sum_{i=1}^{n} \lambda_i}{\mathrm{Tr}(\Lambda)} > \alpha. \tag{1.4}$$

Relatively few signals (10-20; less than 10% of recorded neurons) are required to explain most of the variance ($\geq 75\%$) in motor cortex during point-to-point wrist or arm movements[26,32,35]. Similar insights were provided by Sadtler and colleagues, who trained monkeys to perform point-to-point cursor movements using a brain-computer interface (BCI)[36]. Sadtler *et al.* recorded activity in motor cortex as monkeys passively observed automatic cursor movements, then used factor analysis to map neural activity to a 10-dimensional 'intrinsic manifold' that monkeys quickly learned to use to perform the task, absent any hand movements. The first 4 orthonormalized factors captured roughly 75% of the variance in neural activity – similar to previously discussed findings in non-BCI studies. However, Sadtler *et al.* also provided some of the first insights into how well preserved cortical modes might be across different behaviors. The central question of their study was whether perturbations of the BCI that remained within the intrinsic manifold were easier to learn than perturbations off the intrinsic manifold. Within-manifold perturbations changed the mapping between factors and cursor kinematics (preserving the relationship between neural activity and fac-

tors), while off-manifold perturbations changed the mapping between neural activity and factors (preserving the relationship between factors and cursors kinematics). Monkeys quickly adapted to within-manifold perturbations, but failed to learn off-manifold perturbations, even after 1000 attempts ("trials") to move the cursor from one point to another[36]. These findings argue that it is not only possible to reuse existing neural modes to generate new behaviors, but that learning new modes requires extensive synaptic rewiring, which may be computationally disadvantageous or cumbersome for the central nervous system. To date, only one study has directly investigated whether neural modes are preserved across different behaviors. Gallego and colleagues found that a 12-dimensional manifold identified during one type of wrist movements explained nearly all of the variance in neural activity during different wrist movements[35]. These results have been taken to suggest that motor cortex activity modulates along a low-dimensional manifold to generate different behaviors[6].

It remains unclear whether low-dimensional cortical manifolds for motor control reflect simplicity in cortical processing or simplicity in the tasks typically employed to study motor cortex. Shedding light on this matter would have profound implications for our understanding of motor cortex, future experimental designs, or both. If low-dimensional manifolds reflect cortical simplicity, then that would suggest an exceptionally high degree of redundancy among the 1.3 billion neurons in human M1. Conversely, if low-dimensional manifolds reflect behavioral simplicity, then that would argue that new tasks will be needed to better understand how motor cortex flexibly drives different behaviors. Gao and Ganguli recently explored these matters, observing from a literature review that neural manifolds with dimensionality of roughly 10% of recorded neurons have not only been reported in motor cortex, but also in olfactory, prefrontal, somatosensory, visual, hippocampal, and brainstem areas[37]. To develop a general framework for considering neural dimensionality, the authors note that a neural manifold is an embedding of a task parameter manifold in neural activity space (**fig. 1.2**). Intuitively, the dimensionality of neural activity is fundamentally limited by the dimensionality of the set of stimuli or behaviors employed by experimenters

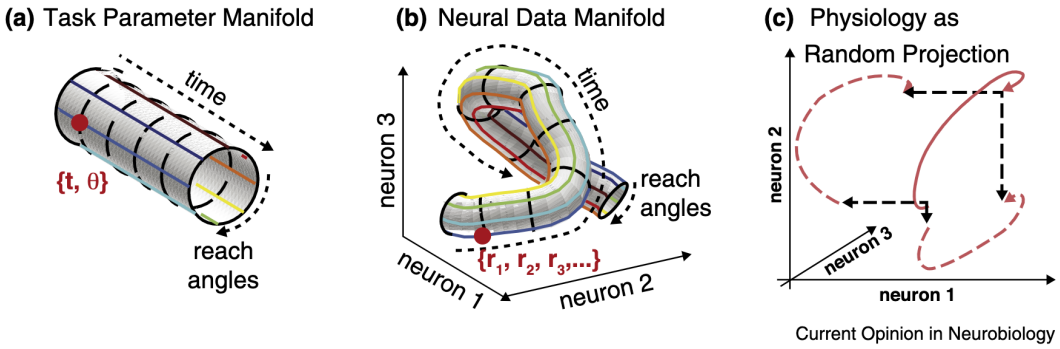


Figure 1.2: **(a)** For a monkey reaching to different directions, the trial averaged behavioral states visited by the arm throughout the experiment are parameterized by a cylinder with two coordinates, reach angle θ, and time into the reach t. **(b)** Trial averaged neural data is an embedding of the task manifold into firing rate space. The number of dimensions explored by the neural data manifold is limited by its volume and its curvature (but not the total number of neurons in the motor cortex), with smoother embeddings exploring fewer dimensions. The [neuronal task complexity] NTC is a mathematically precise upper bound on the number of dimensions of the neural data manifold given the volume of the task parameter manifold and a smoothness constraint on the embedding. **(c)** If the neural data manifold is low dimensional and randomly oriented w.r.t. single neuron axes, then its shadow onto a subset of recorded neurons will preserve its geometric structure. We have shown, using random projection theory [38,39,40] that to preserve neural data manifold geometries with fractional error ϵ, one needs to record $M \geq (1/\epsilon)K \log(\text{NTC})$ neurons. The figure illustrates a $K = 1$ dimensional neural manifold in $N = 3$ neurons, and we only record $M = 2$ neurons. Thus, fortunately, the intrinsic complexity of the neural data manifold (small), not the number of neurons in the circuit (large) determines how many neurons we need to record. (Reproduced from Gao and Ganguli, 2015, with permission).

to evoke neural activity; the number of recorded neurons may or may not reach that fundamental task-imposed limit. Gao *et al.* make this concept concrete by defining a neuronal task complexity (NTC) measure, which captures the maximum number of dimensions neural activity could explore given the constraints of the tasks[38]. The NTC depends on the range of task parameters (e.g., reach duration and angles) and the autocorrelation length of neural activity with each parameter (i.e., how quickly neural activity can change with each parameter). Using data collected by Yu *et al.* from two monkeys during reaching[39], Gao *et al.* compute the NTC as approximately 10, closely matching the original estimate[38]. From this finding, the authors conclude, "if we were to record more neurons, even roughly all 500 million neurons in macaque motor cortex, the dimensionality of the neural manifold in each monkey would not exceed 10.5 and 8.6 respectively[38]." These results argue that the prevailing low-dimensional view of motor cortex may simply reflect the narrow

range of behaviors that have been studied. Given that most studies of motor cortex employed brief, point-to-point reaching paradigms[15–19,24,26–30], considering new tasks may push the boundary on our understanding of the cortical control of movement.

1.4 Motor units and their inputs

If motor cortex is the coach – observing the team from the sidelines, planning their next move and calling the plays – then motor units are the players on the field, doing the physical work of bringing each play to life. A limb muscle is composed of hundreds of thousands of fibers, each containing the contractile machinery needed to shorten the fiber and generate tension in the muscle[7]. Yet the anatomy of the neuromuscular system does not permit each fiber to contract independently. Multiple fibers within a muscle are innervated by a single α-motoneuron, located in the ventral root of the spinal cord[3]. Synaptic transmission from motoneuron to its fibers essentially never fails in humans[40]. Thus, the motoneuron and its fibers operate in tandem; an action potential emitted by a motoneuron causes all of its fibers to concurrently contract. The motoneuron, its axon, and the muscle fibers it innervates are collectively called a motor unit (MU) – the smallest functional element of the neuromuscular system[3]. The motoneurons that collectively innervate all of the fibers within a muscle are distributed along vertical columns in the spinal cord called motor neuron pools[41].

MUs receive input from multiple descending pathways, originating from the brainstem and cerebral cortex, as comprehensively reviewed by Lemon[42]. Brainstem pathways include the tecto-, reticulo-, vestibulo-, and rubrospinal tracts, which largely mediate bilateral postural control and some flexion-based movements of distal limbs, such as the elbow and wrist. Cortical pathways include the corticobulbar tract, which provides input to the brainstem, and the corticospinal tract. The corticospinal tract fulfills multiple functions, including modulating sensory afferents; autonomic control; trophic functions; gain control of spinal reflexes; long-term plasticity of spinal circuits; and excitation and inhibition of MUs[42]. M1 provides input to MUs through the corticospinal tract,

with most of its projections arriving polysynaptically (i.e., routing through one or more spinal interneurons). In humans and some primates, a small proportion of corticospinal neurons constitute the corticomotoneural (CM) tract and synapse monosynaptically onto MUs controlling muscles of the shoulder, elbow, and finger[43]. The functional significance of the CM system remains unclear, but is thought to facilitate fine motor control[42].

In addition to descending inputs, MUs receive a morass of proprioceptive and spinal inputs. Windhorst motivates his comprehensive summary of spinal networks with the following ominous statement from Gerald Loeb:

> Those who believed the spinal cord and peripheral motor plant to be well-understood and thus turned their attentions to higher centers of motor planning and coordination (e.g., cerebral cortex and cerebellum) now find that their edifices are built upon 'the shifting sands of spinal segmental circuitry'. (Windhorst, 2007)[44]

α-motoneurons receive mono- and polysynaptic inputs from each group (I-IV) of sensory afferents (**fig. 1.3**). Ia afferents innervate muscle spindles, discharging in response to changes in muscle length, and provide monosynaptic feedback to α-motoneurons through the commonly studied stretch reflex circuit. γ-motoneurons modulate the sensitivity of Ia discharge. Ib afferents innervate Golgi tendon organs, located at the musculo-tendinous junction, and discharge in response to changes in muscle tension, providing polysynaptic input to α-motoneurons via Ib interneurons. Renshaw cells recurrently inhibit α- and γ-motoneurons. Presynaptic inhibition, mediated by separate GABAergic interneurons, modulates the synaptic efficacy of Ia and Ib afferents. Fatiguing muscle fibers evoke reflex effects from groups III and IV afferents, which modulate α- and γ-motoneurons, Ib interneurons, Renshaw cells, and presynaptic inhibition. These numerous input sources contribute to diversity among MUs (**section 1.5**) and also depend on behavioral context. As Windhorst describes, stretch reflexes were originally characterized by Sherrington using decerebrate preparations, wherein the brainstem in anesthetized animals is surgically transected, which disconnects the brain from the spinal cord while enhancing the drive to extensor α-motoneurons[44].

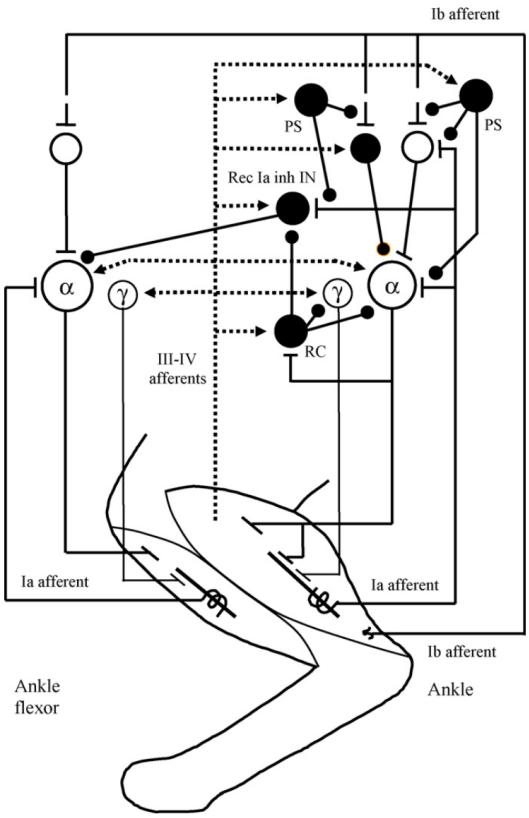


Figure 1.3: Simplified diagram of influences exerted by groups III–IV afferents from extensor muscles on spinal moto- and interneurons. Each neural element represents a population. Excitatory neurons are symbolized by open circles and their synapses by T junctions. At the bottom, a hindleg is sketched with outlines of the ankle flexor muscles (left) and ankle extensor muscle (right). The muscles contain muscle spindles symbolized as straight lines with coils (primary sensory endings) around their middle portions. Spindles lie in parallel to the main skeletal muscle fibers. They receive a motor innervation from α-motoneurons and from branches of β-motoneurons (here called α-motoneurons, see footnote 2). Group Ia afferents originate from primary endings on muscle spindles and project to the spinal cord, in which they make monosynaptic excitatory connections to α-motoneurons of their own (homonymous) muscle and of synergistic muscles (Section9.4.4). Also included are some pathways from extensor group Ib afferents from Golgi tendon organs, which during rest inhibit extensor α-motoneurons and facilitate flexor α-motoneurons (via inhibitory and excitatory interneurons, respectively), while during the stance phase facilitating extensor α-motoneurons via excitatory interneurons, which also in part receive convergent group Ia afferent inputs (for details see text). For simplicity, spindle group II afferents have been omitted. Furthermore, interneurons mediating presynaptic inhibition of sensory afferents are indicated by filled circles denoted "PS". Group III–IV afferents are symbolized by black dotted arrowed lines and may have oligo- and polysynaptic, excitatory or inhibitory effects (for details see text). Abbreviations: PS, interneurons mediating presynaptic inhibition; RC, Renshaw cell. (Reproduced from Windhorst, 2007, with permission)

In decerebrate preparations, extensor Ib afferents inhibit extensor α-motoneurons and excite flexor α-motoneurons; yet the reverse action also occurs during locomotion, indicating that alternative reflex pathways may be involved in different motor tasks. Ia afferents also appear to have differential effects in intact animals, with some evidence suggesting that Ia feedback is significantly reduced during walking but greatly enhanced during sprinting. The discharge of γ-motoneurons, Renshaw cells, and presynaptic inhibition are also all normally modulated by descending commands, including those arriving via the corticospinal tract[44].

1.5 Motor unit diversity

There exists considerable diversity in the morphology, physiology, and inputs to MUs. The most well studied source of diversity is in size. MU size can, somewhat confusingly, refer to multiple quantities: the cell size (diameter) of the α-motoneuron, the diameter of its axon, the number of fibers it innervates, or the diameter of its fibers. By all accounts, these quantities scale proportionally with one another[45]. But conventionally, MU size refers to the number of muscle fibers innervated by the motoneuron, commonly called its 'innervation ratio'[46]. Innervation ratios vary exponentially over a 100-fold range within a motor neuron pool, such that most MUs have small innervation ratios[46]. Physiologically, the innervation ratio principally determines a MU's maximal force capacity[47]. To clarify terminology, an impulse emitted by an α-motoneuron drives all of its muscle fibers to concurrently contract, which generates a brief rise and fall in muscle tension with a stereotyped temporal profile. This muscle-tension response to one MU impulse is called its twitch response[48]. Repeated, high-frequency discharge causes individual MU twitches to fuse into a 'tetanic' response. The force generated by the tetanic response grows linearly at low discharge frequencies, then sigmoidally at higher frequencies[49]. The saturation point in this force-frequency curve is the peak tension, or maximal force capacity of the MU. Like measures of size, maximal force capacity and the amplitude of the twitch response ('peak twitch') scale proportionally with each other and are essentially interchangeable; if the peak twitch for one MU is larger than that of another, then its maximal force capacity will also be larger[49]. Thus, MUs differ in how many

muscle fibers they control, which determines the maximum force generated by a single impulse or sustained discharge.

MUs differ not only in size, but also in speed, due to the mechanical properties of their fibers. One measure of MU speed is its contraction time: the latency between an impulse and the twitch response reaching its peak[50]. Contraction times vary over a 7-10-fold range[50,51], and larger, more forceful MUs tend to control faster contracting muscle fibers[51]. Intriguingly, whereas it was once thought that fiber types are fixed, extensive work demonstrates that muscle fibers are exquisitely plastic, capable of undergoing dramatic, type-specific transitions, as reviewed by Pette[52]. All muscle fibers are uniformly slow contracting in early development[3]. Initial diversification is driven by innervation, and muscle fibers innervated by the same motoneuron are generally homogeneous[53]. Buller first showed that MU types are not immutable, even well after development; switching the nerves that supply slow and fast muscle fibers causes the originally slow fibers to contract more quickly and the fast fibers to contract more slowly[54]. Fiber type transitions due to cross-reinnervation are now understood to be driven by the pattern of motoneuron activation[52]. Fast fibers can be converted into slow fibers through chronic, low-frequency muscle stimulation (mimicking the typical tonic pattern of slow MU activity); slow-to-fast transitions are harder to evoke, but can be achieved with high-frequency stimulation following denervation or nerve blocks[55]. Fiber type transitions are not limited to invasive manipulations. Endurance training increases the prevalence of slow fibers and some evidence indicates that high-intensity sprint training increases the prevalence of fast fiber types[56]. Thus, muscle fiber contractile properties vary broadly across MUs and are adaptable to external stimuli and functional demands.

The distribution of input sources affords one final form of MU diversity to be considered. As briefly aforementioned (**section 1.4**), Renshaw cells mediate recurrent inhibition between α-motoneurons. In fact, the strength of recurrent inhibition varies widely between motoneurons. By intracellularly recording from type-identified motoneurons following nerve stimulation, Friedman

found that recurrent inhibitory postsynaptic potentials are 5-fold larger in slow-twitch (S) than in fast-twitch, fast-fatiguing (FF) MUs[57]. From these results, he surmised that

> ... as more FF units are recruited, as in forceful phasic movements, S unit activity is suppressed by recurrent inhibition from the FF units. Such a relationship would appear to be advantageous under certain conditions, since S motor units may require more than 100 ms to reach peak tension and, hence, could compromise vigorous phasic movement. (Friedman, 1981)[57]

Slow MUs are also known to receive stronger input from Ia afferents via the stretch reflex[44]. Thus, the spinal circuitry supports diversification of inputs into MUs of different sizes and speeds.

1.6 Orderly motor unit recruitment

The nervous system modulates the force generated by a muscle in two ways: MU recruitment and rate coding. Recruitment refers to exciting a MU above its critical firing threshold, into a state of active force production, and rate coding refers to modulating its discharge rate[49]. Based on the preceding discussion in **sections 1.4** and **1.5** – the multitude of descending, intraspinal, and feedback pathways that converge onto MUs; the broad range in MU size and force capacity; the diversity and plasticity in muscle fiber types; and the non-uniform distribution of input strengths mediated by different pathways – one might assume that the nervous system leverages this diversity to control MUs in a flexible and task-specific manner. For example, perhaps it preferentially recruits fast MUs and/or suppresses the discharge of slow MUs to perform fast behaviors, as suggested by Friedman[57]. This assumption might be further bolstered by the realization that the corticospinal tract alone contains 1.1 million axons[42], outnumbering the 60,000 MUs dedicated to limb muscles[58] by a ratio of 18 to 1. Granted, it can be difficult to gauge the functional significance of an input pathway based on a single statistic. In the case of the corticospinal tract, it does fulfill multiple roles and contains a relatively small proportion of CM cells that make monosynaptic connections with MUs[42]. On the other hand, the maximal CM post-synaptic potentials evoked by

stimulating cortex in baboons are larger than the post-synaptic potentials from Ia afferents; and the effectiveness of CM synapses increases with stimulation frequency, which does not occur for Ia afferents[59]. Nevertheless, the canonical understanding of MU control largely disregards much of the diversity among MUs and the functional capacity for motor cortex to leverage that diversity.

The canonical description of MU control rests upon Henneman's size principle[7], though the seeds of that description were sowed well before Henneman. Due to the reliable transmission from α-motoneuron to its fibers (**section 1.4**), muscles essentially act as biophysical amplifiers of the spinal cord; MU impulses can be identified as time-localized waveforms in electromyographic (EMG) signals recorded from the muscle[60]. In 1929, Denny-Brown and Phil found that they could recruit MUs, one at a time, by gradually stretching the soleus of a decerebrate cat (effectively exciting MUs through the monosynaptic stretch reflex)[61]. They further found that releasing the stretch de-recruited MUs in the reverse order that they were recruited, indicating the existence of a well-preserved recruitment order. Interestingly, stretching or releasing the muscle affected MU recruitment, but did not modulate the firing rates of recruited MUs[61]. In the same year, Adrian and Bronk recorded EMG signals during voluntary muscle contractions in humans[62]. They observed that increasing contraction strength both recruited new MUs and increased all of their firing rates.

Several decades later, Henneman provided a tentative explanation for the sequence of MU recruitment. While recording from α-motoneuron axons in isolated ventral root filaments of a decerebrate cat, he electrically stimulated its ipsilateral sciatic nerve and found that the amplitude of newly recruited motoneurons' action potentials increased with the stimulation current[63]. In a landmark 1965 study, Henneman extensively investigated the link between MU size and recruitment order through the use of the stretch reflex[45]. He found that progressively stretching and then releasing the triceps surae in decerebrate cats recruited motoneurons in increasing order by the size of their action potential, then de-recruited motoneurons in the reverse order (**fig. 1.4**). This ordering was not strictly inviolate, but the correlation between size and recruitment order was statistically

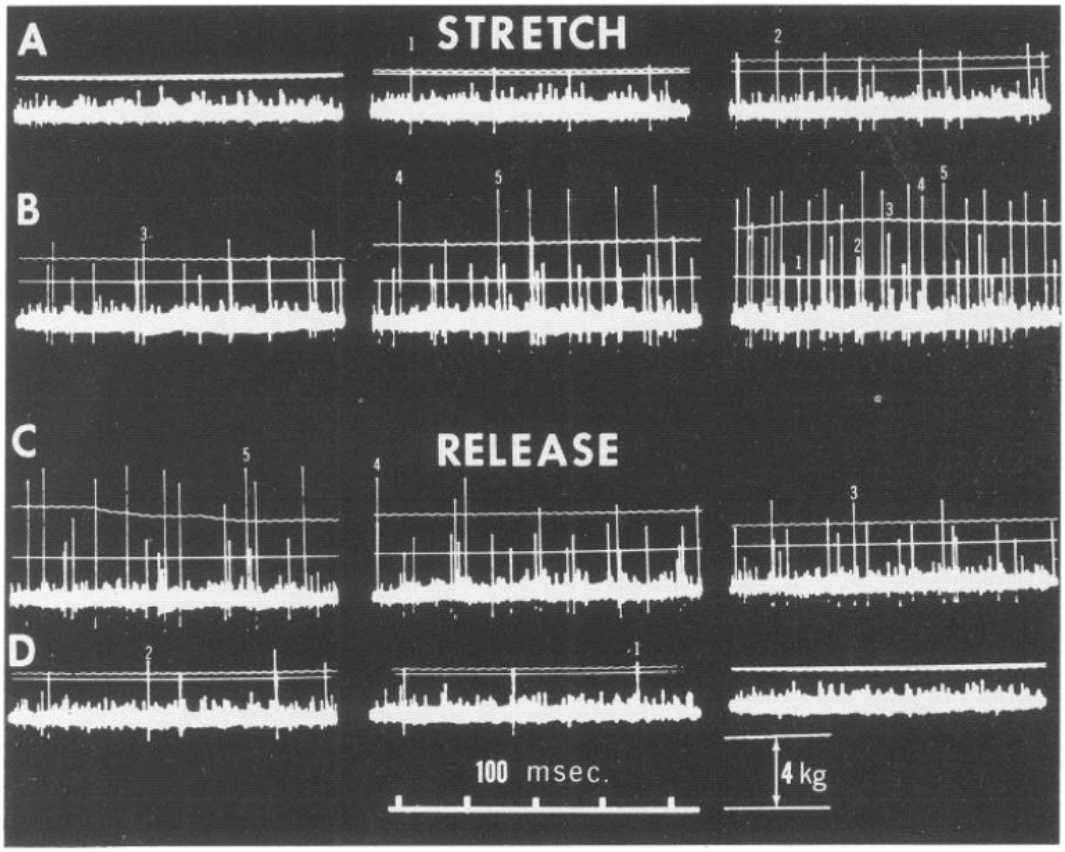


Figure 1.4: Stretch-evoked responses of five alpha motoneurons recorded from a filament of the first sacral ventral root. Small numerals above action potentials indicate rank of units according to size. (Reproduced from Henneman, 1965, with permission)

significant and he reasoned that any deviations were merely due to experimental errors, such as damaging the isolated filaments. Thus, he wrote: "Translating impulse size into fiber diameter and fiber diameter into cell size, we may conclude that there is a general rule or principle applying specifically to motoneurons and perhaps to all neurons, according to which the size of a cell determines its threshold[45]." In subsequent studies, Henneman and colleagues found that multiple spinal reflex pathways[64,65], nerve[66] and supraspinal[67] stimulation in cats all recruited motoneurons in accordance with a 'size principle', wherein small motoneurons are recruited before larger ones. Henneman interpreted these findings as reflecting a computationally advantageous strategy for the nervous system[68]. Since there are practically infinitely many ways that hundreds of MUs could be combined to generate a particular output, dramatically reducing the number of degrees of freedom – to one – would greatly reduce the burden on a central controller. Therefore, Henneman argued

that the size principle represented an all-purpose solution:

> It has been suggested that the various inputs converging on a motoneuron pool may be capable of activating it differentially. It seems unlikely, however, that nature would evolve a highly ordered pool, organized as a fixed hierarchy, with all its related properties, just for reflex use. The formidable problems of selection and combination already noted exist for all types of input and would probably compel the CNS to use the same general solution in all cases. It is not obvious, moreover, that a more flexible, selective system would offer any advantages for control of muscles. The necessity for firing all smaller, more slowly contracting motor units along with the larger, faster ones, as the fixed rank order demands, may appear to be a limitation, but does not, in our opinion, introduce any serious difficulties in control or in economy of operation. The available evidence from this laboratory suggests that the motoneuron pool is used in the same way by all inputs. (Henneman, 1974)[68]

Henneman and colleagues leveraged isolated animal preparations to directly relate the size of a motoneuron to its susceptibility to discharge. Yet the size principle makes specific predictions about the coordinated activation of MUs that can be readily tested even without measurements of cell size.

Fundamentally, Henneman's size principle predicts that MU recruitment is a rigid, orderly process: MUs are activated and deactivated in one consistent order, regardless of the type of input provided to the motoneuron pool. Orderliness in MU recruitment during voluntary contractions was first reported by Adrian and Bronk, but they did not conduct a detailed analysis of the phenomenon[62]. In 1973, Milner-Brown and colleagues provided the first direct evidence that MU recruitment is an orderly process during voluntary isometric contractions of the first dorsal interosseous (a small hand muscle) in humans[69]. Specifically, they found that MU peak twitch forces are strongly positively correlated with recruitment thresholds (i.e., the force level at which a MU begins discharging). Since MU maximal force capacity and size are themselves positively

correlated[47], the finding of Milner-Brown *et al.* indicated that the applicability of the size principle was not limited to isolated preparations of large, postural muscles in cats.

The size principle describes the order in which MUs are recruited/de-recruited to increase/decrease muscle force, but does not describe how MU discharge rates are modulated with changes in muscle force. The first insights were provided by De Luca and colleagues, who analyzed MU firing rates as humans gradually increased and decreased the strength of voluntary contractions over several seconds[70]. They found that the firing rates of all MUs belonging to one muscle increase and decrease together with concurrent changes in muscle force and are also highly correlated on small timescales. This observation was interpreted as indicating that MUs within a motor neuron pool are driven by a common input, which specifies the amplitude of the desired muscle force[71].

Orderly MU recruitment and the hypothesis of a "common drive" are two sides of the same coin. Orderly recruitment describes the typical behavior of MUs, either due to artificial inputs[45,63–66] or during voluntary contractions[69], whereas common drive describes the control scheme used by the nervous system to modulate MU activity[70]. Multiple followup studies spanning several decades corroborated the predictions of orderly MU recruitment and common drive during steady voluntary contractions[51,72–78]. Taken together, these two principles were considered to relieve supraspinal centers from the burden of regulating the activity of individual MUs separately[7,71,79]. Instead, cortex could simply provide a one-dimensional input to a motor neuron pool and passively rely on the physiological differences between MUs (i.e., their size) to determine their recruitment and discharge rate for generating the desired level of muscle force.

1.7 Apparent violations of orderly MU recruitment

The limitations of the size principle have been actively contested for decades. Broadly speaking, violations of orderly recruitment have been studied in two contexts: volitional MU control and task- or context-dependent variations in MU recruitment during natural behavior.

Several studies report that humans can learn to voluntarily control individual MUs. In 1962, Harrison and Mortensen pioneered investigations of volitional MU control in the tibialis anterior[80]. When provided auditory and/or visual feedback of EMG signals, subjects quickly learned to modulate the overall firing rate of individual MUs and even produce precisely timed single, double, triple or quadruple discharge patterns. Some subjects were also able to independently recruit and control up to six different MUs[80]. Similarly fine-grained MU control has been observed in muscles of the hand[81–84] and in the biceps[85,86]. These findings directly conflict with the hypothesis that MU recruitment is an orderly process[51,69–78] strictly enforced by the spinal circuitry[45,65,67], suggesting instead that MU recruitment can be altered at will. An important methodological detail shared by these conflicting bodies of work is that subjects always performed steady, low-intensity muscle contractions, usually isometrically (i.e., preserving muscle length by preventing limb movement with a molded cast or brace).

By the mid 80s, there was growing interest in whether the typical orderliness of MU control merely reflected the narrow range of behaviors typically employed in laboratory settings. Nardone first reported the recruitment of fast-twitch MUs and concurrent de-recruitment of slow-twitch MUs in the gastrocnemii during rapid dorsiflexion in humans[87]. Similar recruitment patterns were reported during running in rats[88,89] and cycling in humans[90]. These studies suggest that the manner in which MUs are controlled during steady isometric contractions differs during dynamic tasks. Other studies considered variations in MU recruitment based on the type of movement or direction of exerted forces. ter Haar Romeny and colleagues found that MU recruitment in the long head of the biceps depends on the motor action; some MUs are recruited exclusively during elbow flexion or forearm supination, while other MUs are recruited based on the linear (or nonlinear) combination of flexion/supination forces[91]. Herrmann and Flanders found that MUs in the biceps and deltoid are "directionally tuned". In subjects generating forces in a 3-dimensional workspace, MUs are preferentially recruited for one or two different force directions[92]. These findings all

seem to refute the notion of a fixed, rigid recruitment order. However, some important technical and scientific caveats have prevented much of this work from substantially altering the canonical view of MU control.

The most direct way to measure MU activity requires inferring MU spike times. Each MU detected by EMG electrodes generates a unique spatiotemporal waveform in the EMG signals every time the MU emits an action potential ("spike")[60]. Inferring all of the unique action potential waveforms contained within the EMG signals and the times that they occur therefore provides precise information about the neural discharge of individual MUs. Increasing the intensity of muscle contractions tends to recruit new MUs and increase the firing rate of all recruited MUs, which substantially increases the chances that two MUs will discharge coincidentally. Coincidental discharge of one or more MUs causes their signature waveforms to overlap on the EMG signals, which complicates identifying each individual MU[60]. Consequently, directly measuring MU activity during dynamic movements is extremely difficult. Most studies of MU recruitment during dynamic tasks circumvented this issue, instead relying on indirect measures of MU activity. Nardone compared individual MU firing rates with the overall activity of the muscle and could not resolve instances of overlapping MU waveforms[87]. More recent investigations[88–90] leveraged an entirely different method of inferring MU activity based on time-frequency decompositions of EMG signals[93]. This method does not provide direct measurements of MU spike times, but affords some information about the relative activation of slow- and fast-twitch MUs[94,95]. Yet the extent to which MU recruitment during natural behaviors can be reliably studied based on spectral properties of EMG signals alone is controversial[96]. Aside from these technical matters, scientific caveats have also complicated the interpretation of prior work suggesting flexibility in MU recruitment.

As originally conceived, the size principle applies to an anatomically defined motor neuron pool (i.e., those motoneurons which innervate all the fibers belonging to one muscle)[45,97]. Yet there are known cases in which one muscle is functionally segregated into "task groups"[98]. In general,

muscles are used in one of three kinematic contexts – isometric (preserving muscle length), concentric (shortening muscle), and eccentric (lengthening muscle) – which, due to the various nonlinear properties of muscle fibers and sensory organs, represent very different operating regimes. For example, muscle spindles provide strong afferent input to α-motoneurons during isometric and lengthening contractions, but spindle afferents turn off with high rates of muscle shortening[98]. Thus, Loeb proposed: "When a muscle is required to perform more than one kinematic type of task (e.g. active lengthening vs active shortening), it may solve the control circuit problem by functionally dividing itself into two separate task-orientated groups.[98]" The action of the sartorius muscle during walking is a prime example. Sartorius is a biarticular muscle that is divided into two anatomically distinct compartments (corresponding to two different motor neuron pools); the anterior head flexes the knee and the medial head extends the knee[99]. Yet *within the anterior head itself*, there are two distinct groups of motoneurons, one which is active during the stance phase of a step cycle and the other is active during the swing phase[99]. This finding would conflict with the size principle under a strict anatomical definition of a motor neuron pool but not under a refined definition that accounts for task groups in multifunctional muscles:

> The simplifying "size principle" apparent in the inputs to and the recruitment of the motor pools of unifunctional muscles could still hold for the motoneurons comprising each group. The traditional concept of "motor pool," which stemmed from purely anatomical considerations, would thus be replaced with the concept of task-specific groups of motoneurons whose properties would be defined on functional, in addition to morphological, criteria. (Hoffer *et al.*, 1987)[99]

It has therefore been argued that apparent violations of the size principle be evaluated under an "operational" definition of a motor neuron pool[100]. Much of the work demonstrating task-dependent[91,92] or volitional[83,85,86] MU control were conducted in multifunctional muscles. In some cases, the authors explicitly noted that volitional MU control was only achievable in multifunctional muscles (but not in unifunctional muscles)[84] or when subjects were allowed to change postures[82,85]. Thus, it remains unclear whether meaningful violations of orderly recruitment ac-

tually occur or merely reflect the known properties of multifunctional muscles[100]. Furthermore, analyses of the mechanical properties of muscles in response to different artificially induced patterns of MU recruitment indicates that selectively recruiting fast-twitch MUs may not actually confer any functional advantage[101]. It also remains controversial whether fast-twitch MUs are, in fact, selectively recruited during rapid voluntary contractions[102]. Thus, as described in C.J. Heckman's comprehensive survey of motor units, despite "some suggestions that certain conditions require some flexibility [in motor unit recruitment][103]", the general consensus remains that MU recruitment during voluntary contractions adheres to the predictions of Henneman's size principle.

Chapter 2: Flexible Neural Control of Motor Units

Voluntary movement requires communication from cortex to the spinal cord, where a dedicated pool of motor units (MUs) activates each muscle. The canonical description of MU function, established decades ago, rests upon two foundational tenets. First, cortex cannot control MUs independently[7] but supplies each pool with a common drive that specifies force amplitude[70,71]. Second, as force rises, MUs are recruited in a consistent order[51,62,69,72–78] typically described by Henneman's size principle[45,63,64,66,68]. While this paradigm has considerable empirical support, a direct test requires simultaneous observations of many MUs over a range of behaviors. We developed an isometric task that allowed stable MU recordings during rapidly changing force production. MU responses were surprisingly flexible and behavior-dependent. MU activity could not be accurately described as reflecting common drive, even when fit with highly expressive latent factor models. Neuropixels probe recordings revealed that, consistent with the requirements of fully flexible control, the cortical population response displays a surprisingly large number of degrees of freedom. Furthermore, MUs were differentially recruited by microstimulation at neighboring cortical sites. Thus, MU activities are flexibly controlled to meet task demands, and cortex has the capacity to contribute to that ability.

2.1 Introduction

Primates produce myriad behaviors, from acrobatic maneuvers to object manipulation, all requiring precise neural control of muscles. Each muscle is controlled by a motor neuron pool containing hundreds of anatomically and functionally diverse motor units (MUs)[4]. One MU is defined as a spinal α-motoneuron and the muscle fibers it uniquely innervates[3]. MUs are highly heterogeneous[104], differing in size (large MUs innervate more fibers), duration of generated force[104], and the muscle length where force is maximal[105].

Optimality suggests using MUs best suited to the specific situation[92]. Yet such flexibility would necessitate non-trivial computational resources, including participation by brain areas aware of the full movement and context. A simpler alternative is a spinally implemented recruitment strategy that approximates optimality in limited contexts. Supported by nearly a century of research, this alternative has become the canonical conception of MU control[7,106]. In decerebrate cats, MUs are recruited and de-recruited in a consistent order[61] from smallest to largest according to Henneman's size principle[45,63,64,66,68]. Orderly MU recruitment is similarly observed following supraspinal stimulation in cats[67] and during voluntary muscle contractions in humans[51,62,69,72–78]. MU firing rates increase monotonically with force and display correlated fluctuations[70], arguing that MUs are jointly controlled by a one-dimensional (1D) 'common drive'[71]. This 'rigid control' hypothesis – common drive followed by small-to-large recruitment – is codified in standard models of muscle activation[49,107].

Rigid control is believed to relieve cortex from the burden of controlling MUs independently[68,71]. Small-to-large recruitment minimizes fluctuations during constant force production and is thus optimal in that context[108–110]. In idealized form, rigid control describes each muscle and its MU pool. There are known exceptions[91,92,98,111,112] when a 'multifunctional' muscle pulls in different directions (necessitating more than one common drive[92,111]) or drives movement across two joints

and participates in multiple synergies[91]. These properties are compatible with rigid control under an operational definition of an MU pool[100,113,114]; descending commands can remain simple (specifying force direction[92] or synergy activation[91]) and small-to-large recruitment holds for any given force direction.

The alternative to rigid control is highly flexible MU recruitment that adapts to situational demands. Some flexibility has been observed during locomotion[88,89], where it may reflect the need to control force when a muscle lengthens under load[87,98,115]. It also seems intuitive that recruitment should favor fast-twitch MUs when forces change rapidly. Yet it remains controversial whether speed does[100,103,116] or should[101] influence recruitment.

Rigid control is thus believed to govern the vast majority of cases[101], with exceptions being rare and/or inconsistent across studies[100,103]. An accepted caveat is that critical tests have yet to be performed[100]. Due to the difficulty of recording many MUs during swiftly changing forces[116], no study has directly addressed the key situation where rigid and flexible control make divergent predictions: when a subject skillfully performs diverse movements, is MU recruitment altered to suit each movement? An additional key test also remains. Fully flexible control would require, in addition to spinal mechanisms, some influence from areas aware of overall context. Flexible control thus makes the strong prediction that altering cortical activity should alter recruitment. This does not occur in cat[67], but remains to be examined in primate.

2.2 Results

2.2.1 Pac-Man Task and EMG recordings

We trained one rhesus macaque to perform an isometric force-tracking task. The monkey modulated force to control the vertical position of a 'Pac-Man' icon and intercept scrolling dots (**fig. 2.1a**). We could request any temporal force profile by appropriately selecting the dot path. We conducted three experiment types, each using dedicated sessions. Dynamic experiments em-

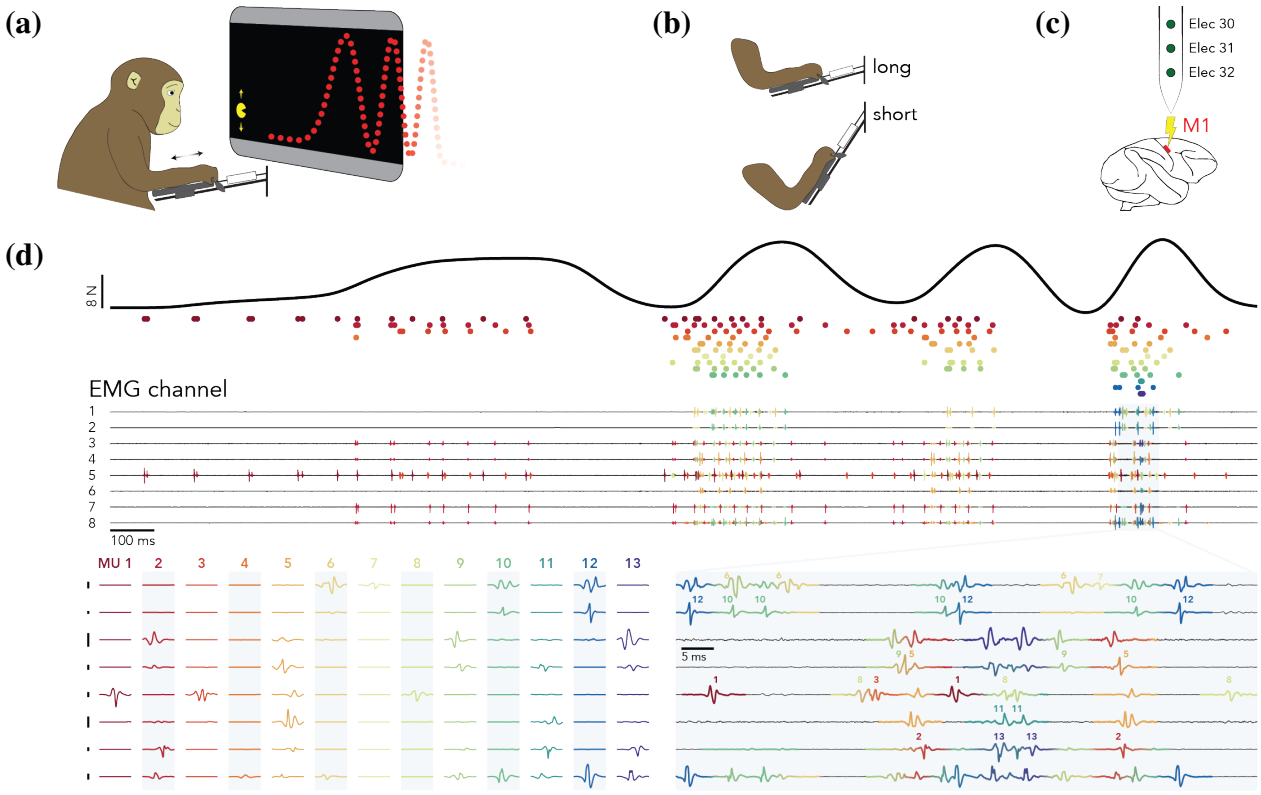


Figure 2.1: Experimental setup and MU spikes. (**a**) Dynamic experiments. A monkey modulated the force generated against a load cell to control Pac-Man's vertical position and intercept a scrolling dot path. A variety of force profiles were used, a subset of which were also employed during muscle-length and microstimulation experiments. (**b**) Muscle-length experiments. The manipulandum was positioned so that the angle of shoulder flexion was 15° (long) or 50° (short). (**c**) Stimulation experiments. Intracortical microstimulation was delivered through a linear array inserted in sulcal motor cortex. (**d**) Behavior and MU responses during one dynamic-experiment trial. The target force profile was a chirp. *Top*: generated force. *Middle*: eight-channel EMG signals recorded from the lateral triceps. 20 MUs were isolated across the full session; 13 MUs were active during the displayed trial. MU spike times are plotted as circles (one row and color per MU) below the force trace. EMG traces are colored by the inferred contribution from each MU (since spikes could overlap, more than one MU could contribute at a time). *Bottom left*: waveform template for each MU (*columns*) and channel (*rows*). Templates are 5 ms long. As shown on an expanded scale (*bottom right*), EMG signals were decomposed into superpositions of individual-MU waveform templates. The use of multiple channels was critical to sorting during challenging moments such as the one illustrated in the expanded scale. For example, MU2, MU5, and MU10 had very different across-channel profiles. This allowed them to be identified when, near the end of the record, their spikes coincided just before the final spike of MU12. The ability to decompose voltages into a sum of waveforms also allowed sorting of two spikes that overlapped on the same channel (e.g., when the first spike of MU6 overlaps with that of MU10, or when the first spike of MU9 overlaps with that of MU5). Multiple channels also guarded against mistakenly sorting one unit as two if the waveform scaled modestly across repeated spikes (as occurred for a modest subset of MUs).

ployed many force profiles including slow and fast ramps and sinusoids (**fig. S2.6**). Muscle-length experiments (**fig. 2.1b**) investigated whether MU recruitment reflects joint angle/muscle length, using a subset of force profiles. Microstimulation experiments (**fig. 2.1c**) artificially perturbed descending commands using microstimulation delivered via a linear electrode array in sulcal primary motor cortex (M1).

On each of 38 sessions, we recorded from multiple custom-modified percutaneous thin-wire electrodes closely clustered within the head of one muscle. Recordings were made from the triceps and deltoid (dynamic experiments); deltoid (muscle-length experiments); and triceps, deltoid, and pectoralis (microstimulation experiments). It is notoriously difficult to spike sort EMG signals during dynamic tasks; movement threatens recording stability and vigorous muscle contractions cause MU action-potential waveforms to superimpose[60]. Three factors enabled us to identify the spikes of multiple single MUs, even during high-frequency force oscillations (**fig. 2.1d**). First, the isometric task facilitated stable recordings even when force changed rapidly. Second, activity intensity could be titrated via the gain linking force to Pac-Man's position. Finally, a given MU typically produced a complex waveform spanning many channels (**fig. 2.1d**, *bottom*), which we identified by adapting recent advances in spike sorting[117–119], including methods for resolving superimposed waveforms[118] (**section 2.6.1**).

We isolated 3-21 MUs in each session (356 total units). Analyses considered only simultaneously recorded neighboring MUs, for two reasons. First, many of our recordings were from the deltoid, where different regions pull in different directions and physically distant MUs thus have different 'preferred force directions'[92]. Second, across-session behavioral variability could conceivably make recruitment order appear inconsistent in pooled data. We thus compared only amongst simultaneously recorded neighboring MUs within a single muscle head, with all forces generated in one direction.

2.2.2 Motor unit activity during behavior

Rigid control applies to the behavior of the full MU pool, yet provides constraints that can be visualized at the level of MU pairs. MUs should be recruited in a consistent order[45,68] and changes in their activity should not be strongly opposing[72,77]. Responses of MU pairs were typically consistent with these predictions during gradually changing forces at a single muscle length. For example, in **fig. 2.2a**, MU88 is recruited before MU90 and the activity of both increases monotonically with increasing force. Both become less active as force decreases, with MU88 de-recruited last. The predictions of rigid control sometimes held during swiftly changing forces (**fig. 2.2b**). However, violations were common when comparing rapidly and slowly changing forces. For example, in **fig. 2.2e**, MU309 is more active during the sinusoid (*left*) than during the ramp (*right*), while the opposite is true for MU311. Thus, which of the two MUs contributes the most to force production depends on context.

Recruitment incompatible with rigid control also occurred within individual conditions if force changed at different rates during different epochs. In **fig. 2.2c**, MU109's activity rises threefold in anticipation of sudden force offset, even as MU108's activity declines. In **fig. 2.2d**, over the last three cycles of a chirp force profile, MU324's activity decreases as MU329's activity increases. These examples are inconsistent with common drive, which cannot simultaneously increase and decrease. For both pairs, the key violation (activity decreasing for one MU while increasing for another) was not observed when holding a static force. Instead, activity reflected whether forces were, or soon would be, rapidly changing. Yet it was rarely the case that MU activity simply reflected the derivative of force. The rate of MU109 (**fig. 2.2c**) rises while force is constant. And while MU329 (**fig. 2.2d**) and MU309 (**fig. 2.2e**) are more active during higher-frequency forces, they do not phase lead their neighboring MUs. Activity reflecting not just force, but the overall situation, was particularly evident with changes in muscle length, both when activity was swiftly changing (**fig. 2.2f**) and when it was static (**fig. 2.2g**). Changes in recruitment with muscle length could be large (e.g., an MU becoming inactive for a given posture) but could also be more modest,

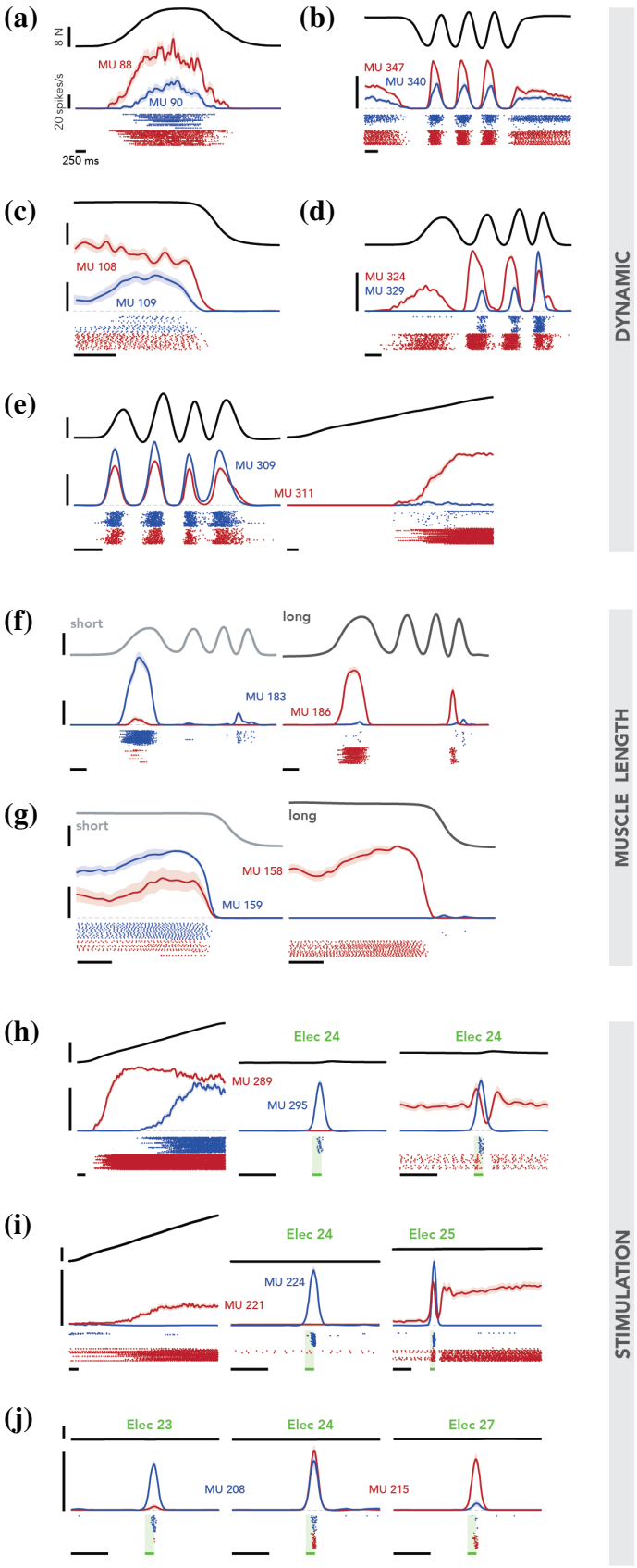


Figure 2.2: Example MU responses. Each panel displays the trial-averaged force (*top*), mean firing rate with standard error (*middle*) and spike rasters (*bottom*) for a pair of concurrently recorded MUs during dynamic (**a**-**e**), muscle-length (**f**,**g**), and stimulation (**h**-**j**) experiments. Vertical scale bars indicate 8 N (forces) and 20 spikes/s (firing rates). Horizontal scale bars indicate 250 ms. Columns within panels correspond to different conditions. In ***h***-***j***, labels indicate the stimulation electrode. Bars (with shaded region overlapping rasters) indicate stimulation duration. On average (across all sessions and experiments) each condition consisted of 34 trials.

allowing us to confirm recording stability (**fig. S2.7**).

2.2.3 Cortical perturbations

Many aspects of the flexibility that we observed (especially those reflecting muscle length) are likely due to spinally implemented flexibility. Yet recruitment reflected factors beyond force and its instantaneous derivative, including whether force would soon change or was overall high frequency, suggesting that supraspinal mechanisms may contribute. If so, it should be possible to alter recruitment by artificially perturbing descending cortical commands[120]. The opposite prediction is made by the classical hypothesis that rigid control is fully enforced at or near the MU pool[45]. If so, perturbation-induced activity, while unnatural in time course, should display orderly recruitment[67]. We manipulated M1 activity using microstimulation (57 ms, 333Hz). Penetration locations and electrode choices were optimized to activate the recorded muscle.

Cortical perturbations often produced unexpected recruitment patterns. For example, given recruitment during a slow force ramp (**fig. 2.2h**, *left*), any common drive that activates MU295 (*blue*) should also activate MU289 (*red*). Yet stimulation on electrode 24 activated MU295 but not MU289 (*center*). When MU289 was already active during static force production, stimulation had an effect consistent neither with common drive (it differed for the two MUs) nor with natural recruitment (where MU289 was always more active). Similarly, in **fig. 2.2i**, stimulation on electrode 24 selectively activated MU224, although MU221 was lower-threshold during a force ramp. Occasionally, cortical perturbations produced hysteresis (**fig. 2.2i**, *right*), likely reflecting persistent inward currents[121]. Unlike the direct effect of stimulation, hysteresis rarely altered recruitment order (activity was higher for MU221, as during natural recruitment)[103].

Thus, in primates, MU recruitment is readily altered by cortical perturbations. Indeed, neighboring MUs, recorded on the same set of closely-spaced electrodes, were often differentially recruited by physically proximal stimulation sites (100 μm electrode spacing). For example, in

fig. 2.2j, electrode 23 recruits MU208, electrode 27 recruits MU215, and electrode 24 recruits both. It remains unclear to what degree the capacity for fine-grained control is typically used (stimulation is an intentionally artificial perturbation), but cortex certainly has the capacity to influence recruitment.

2.2.4 State space predictions of rigid control

The predictions of flexible and rigid control can be evaluated by plotting the activity of two MUs jointly in state space. Under rigid control, MU activity increases nonlinearly but monotonically with force magnitude[49,73,107] (**fig. 2.3a**). Thus, when represented as a point in state space, activity should move farther from the origin with increasing force, tracing a curved (due to the nonlinearities) monotonic one-dimensional (1D) manifold. The manifold is 1D because the activity of every MU varies with a common drive. The manifold is monotonic because, as common drive increases, the rate of every MU increases (or stays the same if unrecruited or at maximal rate). Because each MU has a static link[49,107] function (transforming common drive into a firing rate) manifold shape is preserved across situations. This formulation simply restates the fundamental tenets of rigid control: different MUs are recruited at different times and in different ways, but all have activity that is a monotonic function of a common drive. Thus, under rigid control, the 1D manifold can take any monotonically increasing shape, but activity should always lie on the same manifold (**fig. 2.3b**). In contrast, flexible control predicts that activity will exhibit many patterns that cannot be described by a monotonic 1D manifold (**fig. 2.3c**).

We used the state-space view to examine the joint activity of MU pairs, including many of those from **fig. 2.2**. Examining activity in both formats helps determine whether apparent departures from rigid control are real and nontrivial. Under rigid control, the identity of the most-active MU may reverse if the later-recruited MU has a steeper link function. Without close inspection, this might appear to violate rigid control when plotting activity versus time. In contrast, the state space view would demonstrate that activity remains on a monotonic 1D manifold. Conversely,

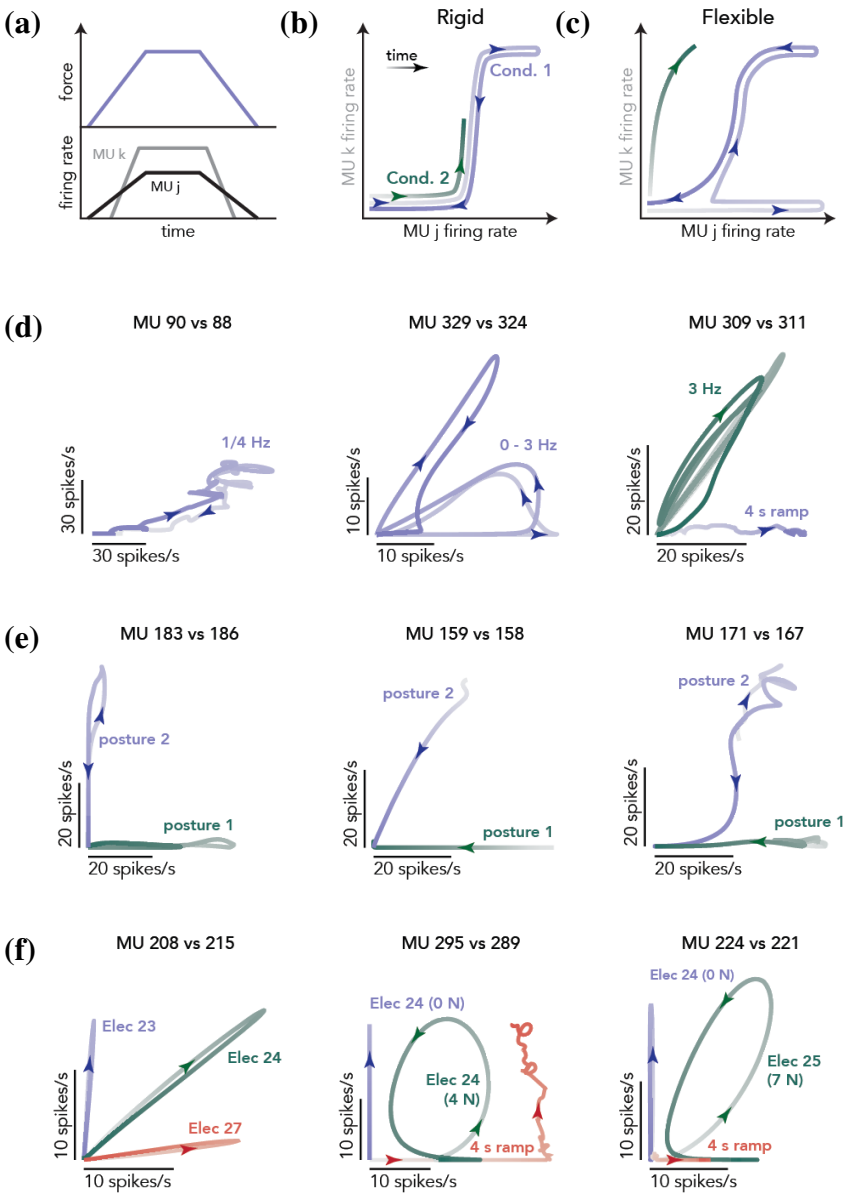


Figure 2.3: State space predictions for rigid and flexible MU control. (**a**) Schematic illustrating firing rates for a pair of hypothetical MUs that are consistent with rigid control. (**b**) Firing rates of the same hypothetical MUs, plotted in state space, for the condition in ***a*** (*purple*) and another idealized condition (*green*). Lines are shaded light-to-dark with the passage of time. Because these hypothetical MUs obey rigid control, activity lies on a 1-D monotonic manifold. (**c**) Schematic of how MU activity could evolve if control is flexible rather than rigid. Activity does not lie on a 1-D monotonic manifold. (**d**) State-space plots for three MU pairs recorded during dynamic experiments. (**e**) State-space plots for three MU pairs recorded during muscle-length experiments for a 250 ms ramp down force profile. (**f**) State-space plots for three MU pairs recorded during microstimulation experiments.

MUs with different latencies could create brief departures from a monotonic 1D manifold, but the lack of a true violation would be apparent when plotting activity versus time.

When considering only slowly changing force profiles, activity typically approximated a monotonic 1D manifold (**fig. 2.3d**, *left*). During rapidly changing forces, activity often deviated from a monotonic 1D manifold either within a condition (**fig. 2.3d**, *center*) or relative to conditions with slowly changing forces (*right*). 'Looping' within a single rapid cycle could reflect latency differences rather than a true violation. However, rigid control is inconsistent with the differently oriented loops across cycles within a chirp force profile (**fig. 2.3d**, *center*) and with the very different trajectories during a 3 Hz sinusoid and a slow ramp (*right*). Large deviations from a monotonic 1D manifold were also observed across muscle lengths (**fig. 2.3e**). Cortical perturbations often drove deviations, both when comparing among electrodes and when comparing with natural recruitment (**fig. 2.3f**).

Quantification across all simultaneously recorded MU pairs confirmed that departures from a monotonic 1D manifold were usually small when considering only slowly changing forces within a single muscle length. Departures were larger when also considering rapidly changing forces, both muscle lengths, or cortical perturbations (**figs. S2.8** and **S2.9**). This effect was seen in 36 of 38 sessions. This quantification was highly conservative; departures were nonzero only if they could not be attributed to latency differences when comparing just two moments of time. To consider how well rigid control describes MU activity across all times and conditions, we leveraged a model-based approach.

2.2.5 Latent factor model

The central tenet of rigid control is that all MUs within a pool are controlled by a common drive; different MU activities arise from MU-specific link functions of that drive. We wished to quantitatively evaluate how well this model can account for the joint activity of all simultaneously

recorded MUs. Conceptually this approach is simple: the model should be rejected if it fits the data poorly even when granted full expressivity (no constraints other than those inherent to rigid control). Existing models of MU control employ idealized link functions (rectified linear[49] or sigmoidal[107]). While reasonable, those choices limit expressivity. We instead employed a probabilistic latent factor model (**fig. 2.4a**) where the rate of each MU is a function of common drive: $r_i(t) \sim f_i(x(t+\tau_i))$. Model fitting used black box variational inference[122] to infer $x(t)$ and learn the MU-specific f_i and time-lag, τ_i. f_i was unconstrained other than being monotonically increasing.

The resulting model obeys rigid control but is otherwise highly expressive; it can assume essentially any common drive and set of link functions. Because MUs can have different latencies, it can produce some departures from a monotonic 1D manifold. The model provided good fits during slowly changing forces (**fig. 2.4b**, *top*). Fit quality suffered in all other situations, including cortical perturbations (**fig. 2.4b**, *bottom*), because the model could not account for the manifold changing flexibly across situations.

For each session, we fit the activity of all MUs during the 4-second increasing ramp condition, either alone or collectively with other conditions. Error was always computed during the 4-second increasing ramp only. This allowed us to ascertain whether the model's ability to account for activity during a 'traditional' situation was compromised when it had to also account for other situations. Fit error was cross-validated (using random data partitions) and thus should be zero on average for an accurate model. That property was confirmed using an artificial MU population that could be described by one latent factor but was otherwise realistic (accomplished by reconstructing each MU's response from the first population-level principal component, followed by a rectifying nonlinearity). Fit error was indeed nearly zero for the artificial population (**fig. 2.4c**, *filled circles*) regardless of how many conditions were fit.

For the empirical data, fit error was nearly zero when the latent variable model was fit only

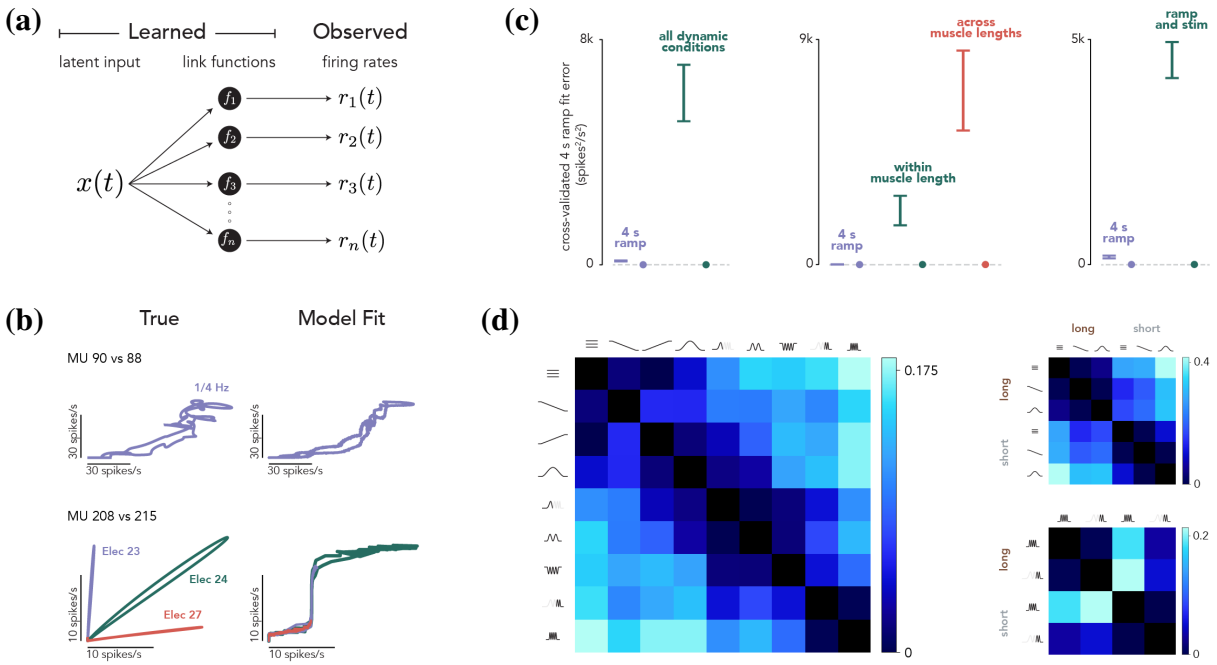


Figure 2.4: Latent Factor Model. (**a**) The premise of rigid control is that MU firing rates are fixed 'link functions' of a shared, 1-D latent input. If so, it should be possible to infer link functions and a latent that account for the observed rates. We assessed the degree to which this was true, with essentially no constraints other than that link functions be monotonically increasing. (**b**) Illustration of model fits for simplified situations: the activity of two MUs during a slow ramp (*top*) or following cortical stimulation on three different electrodes (*bottom*). (**c**) Quantification of model performance when accounting for the activity of the full MU population. Cross-validated error was the median (across MUs) dot product of the model residuals (difference between actual and model MU activity) for random splits of trials during the slow-ramp condition alone. Cross-validated error was computed when the model only had to fit the ramp condition, or also had to fit other conditions. *Left*. For dynamic experiments, the other conditions were the different force profiles. *Center*. For muscle-length experiments, the other conditions were different force profiles using the same muscle length (a subset of the force profiles used in the dynamic experiments) or all force profiles across both muscle lengths. *Right*. For microstimulation experiments, the other conditions involved cortical stimulation (on one of 4-6 electrodes) during static force production at different levels. Error bars indicate the mean ± standard deviation of the cross-validated error across 10 model fits, each using a different random division of the data to compute cross-validated error. Circles indicate fit error when the model was fit to an artificial population response that truly could be described by rigid control, but otherwise closely resembled the empirical population response. (**d**) Proportion of total MUs that consistently violated the 1-latent model when fit to pairs of conditions. Each entry is the difference between the proportion of consistent violators obtained from the data and the proportion expected by chance. *Left*: dynamic experiments; *right*: muscle-length experiments.

during the 4-second increasing ramp (**fig. 2.4c**, *purple*). Fits were compromised when the model had to also account for dynamically changing forces, different muscle lengths, or cortical perturbations. This combination of findings explains why the hypothesis of rigid control was appealing (it can describe responses when forces change slowly) while also demonstrating that it fails to describe MU activity once a broader range of behaviors is considered.

We used a complementary approach, focused on single trials, to further explore when the model of rigid control failed. We fit the model to single-trial responses from two conditions at a time. We defined an MU as a 'consistent violator' if its activity was overestimated for trials from one condition and underestimated for trials from the other condition, at a rate much higher than chance. Consistent violators indicate that recruitment differs across conditions in a manner inconsistent with rigid control. Consistent violators were relatively rare when two conditions had similar frequency content (**fig. 2.4d**, *left, dark entries near diagonal*), but became common when conditions had dissimilar frequency content. Additionally, consistent violators became common when comparing across muscle lengths (**fig. 2.4d**, *right*).

2.2.6 Neural degrees of freedom

If a one-degree-of-freedom (common) drive cannot account for MU activity, how many degrees of freedom must one assume (**fig. 2.5a**)? To identify a lower bound, we fit models with multiple latent factors for two dynamic-experiment sessions (those with the most simultaneously recorded active MUs: 16 and 18). Cross-validated fit error (**fig. 2.5b**) reached zero around 4-6 factors. Thus, describing the activity of the 16-18 MUs required 4-6 degrees of freedom. Because we recorded a minority of MUs (the triceps alone contain a few hundred) from a localized region during a subset of behaviors, there are likely many more degrees of freedom even for a given muscle. Neural control of the arm may thus be quite high-dimensional, with dozens or even hundreds of degrees of freedom once all muscles are considered.

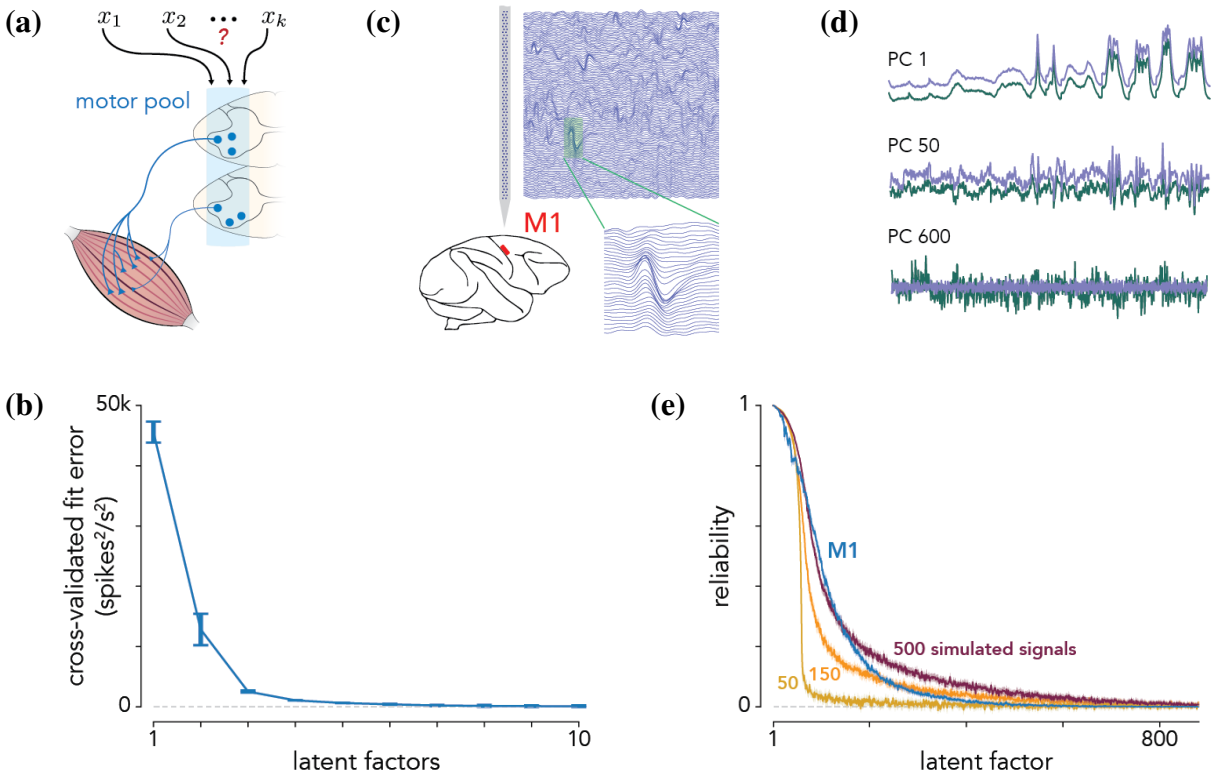


Figure 2.5: Quantifying neural degrees of freedom. (**a**) We considered the number of latent inputs that drive MU activity. (**b**) Cross-validated fit error for models with 1-10 latent factors. Cross-validated error (**fig. 2.4c**) was computed across all dynamic-experiment conditions. Error bars indicate the mean ± standard deviation of that median error across 10 model fits, each using a different random division of the data to compute cross-validated error. (**c**) We recorded neural activity in M1 using 128-channel Neuropixels probes. (**d**) Two sets of trial-averaged firing rates were created from even and odd trials. Traces show the projection of the even (*green*) and odd (*purple*) population activity onto three principal components (PCs) obtained from the even set. Traces for PCs 1 and 50 were manually offset to aid visual comparison. (**e**) Reliability of neural latent factors. Two sets of trial-averaged data were obtained from random partitions of single trials. Both data sets were projected onto the principal components (factors) obtained from one set. The reliability of each factor was computed as the correlation between the projection of each data set onto the factor. Traces indicate the mean and 95% confidence intervals (*shading*) for 25 re-samples of M1 activity (*blue*) and simulated data with 50 (*yellow*), 150 (*orange*), and 500 (*red*) latent signals.

It is unclear how many of these degrees of freedom are influenced by descending control. Anatomy suggests it could be many. The corticospinal tract alone contains approximately one million axons[42], and our perturbation experiments reveal a potential capacity for fine-grained control. A counterargument is that descending commands must be drawn from dimensions occupied by cortical activity, which is typically described as residing in a low-dimensional manifold[25,35,36]. In

standard tasks, 10-20 latent factors account for most of the variance in M1 activity[26,32,35,36,123]. Yet the remaining structure, while small, may be meaningful[22] given that descending commands appear to be small relative to other signals[31].

We reassessed the dimensionality of activity in M1, aided by three features. First, our task involves force profiles spanning a broad frequency range, potentially revealing degrees of freedom not used in other tasks. Second, we considered an unusually large population (881 sulcal neurons) recorded over multiple sessions using the 128-channel version of primate Neuropixels probes (**fig. 2.5c**). Third, to assess whether a latent factor is meaningful, we focused not on its relative size (i.e., amount of neural variance explained) but on whether it was reliable across trials (**fig. 2.5d**) using a method similar to that of Stringer and colleagues[124]. When analyzing a subset of neurons, a small but meaningful signal (e.g., one that could be reliably decoded from all neurons) will be corrupted by spiking variability but will still show some nonzero reliability across trials. We defined reliability, for the projection onto a given principal component, as the correlation between held-out data and the data used to identify the principal component (**fig. 2.5d**).

The first two hundred principal components all had reliability greater than zero (**fig. 2.5e**). To put this finding in context, we analyzed artificial datasets that closely matched the real data but had known dimensionality. Even when endowed with 150 latent factors, artificial populations displayed reliability that fell faster than for the data. This is consistent with the empirical population having more than 150 degrees of freedom, an order of magnitude greater than previously considered[26,32,35,36,123]. For comparison, if M1 simply encoded a force vector, there would be only one degree of freedom; all forces in our experiment were in one direction. Encoding of the force derivative would add only one further degree of freedom. Thus, the M1 population response has enough complexity that it could, in principle, encode a great many outgoing commands beyond force per se. Direct inspection of individual-neuron responses (**fig. S2.10**) supports this view; neurons displayed a great variety of response patterns.

2.3 Discussion

The hypothesis of rigid control – a common drive followed by size-based recruitment – has remained dominant[7,106] for three reasons: it describes activity during steady force production[51,62,69,72–78], would be optimal in that situation[109], and could be implemented via simple mechanisms[45,68]. It has been argued that truly flexible control would be difficult to implement and that "it is not obvious... that a more flexible, selective system would offer any advantages.[68]" Yet there has existed evidence, often using indirect means, for at least some degree of flexibility in specific situations[103]. Our findings argue that flexible MU control is likely a normal aspect of skilled performance in the primate. Recruitment differed anytime two movements involved different force frequencies or muscle lengths.

An appealing hypothesis is that flexibility reflects the goal of optimizing recruitment for each behavior. To test the internal validity of this hypothesis, we employed a normative model of force production by an idealized motor pool where MUs varied in both size and how quickly force peaked and decayed (Supp. Materials). The model employed whatever recruitment strategy maximized accuracy, using knowledge of future changes in force. During slowly changing forces, the model adopted the classic small-to-large recruitment strategy (**fig. S2.12**). During rapidly changing forces, the model adopted different strategies that leveraged heterogeneity in MU temporal force profiles. From this perspective, the size principle emerges as a special case of a broader optimality principle.

Optimal recruitment would require cooperation between spinal and supraspinal mechanisms. Our data support this possibility. Flexibility driven by changes in muscle length presumably depends upon spinally available proprioceptive feedback[114]. During dynamic movements, some aspects of flexibility reflect future changes in force, which would likely require descending signals. The nature of the interplay between spinal and descending contributions remains unclear, as is

the best way to model flexibility. Flexibility could reflect multiple additive drives to the MU pool and/or modulatory inputs that alter input-output relationships[121] (i.e., flexible link functions). Both mechanisms could have spinal and/or supraspinal sources.

The hypothesis that descending signals influence MU recruitment has historically been considered implausible, as control might be unmanageably complex unless degrees of freedom are limited[68,71]. Indeed, descending control has typically been considered to involve muscle synergies[125], without even the ability to independently control individual muscles. Consistent with that view, recruitment order is unaltered by supraspinal stimulation in cats[67]. Recruitment can be altered using biofeedback training in humans[80,81], although it is debated whether this ability reflects unexpected flexibility or simply leverages known compartmentalization of multifunctional muscles[84,126].

In our view there is little reason to doubt the existence of descending influences on MU recruitment. The corticospinal tract alone contains on the order of a million axons[42], including direct connections onto α-motoneurons[42,43] from neurons whose diverse responses[127] reflect the context in which a force is generated[128]. Our findings supply three additional reasons to suspect rich descending control. First, during a learned task performed skillfully, recruitment is far more flexible than previously thought. Second, stimulation of neighboring cortical sites can recruit neighboring MUs, disproving the assumption that "the brain cannot selectively activate specific motor units[7]". Third, M1 activity has a surprisingly large number of degrees of freedom that could potentially contribute descending commands. Future experiments will need to further explore whether MU recruitment is fully or partially flexible, the level of granularity of descending commands, and how those commands interact with spinal computations.

2.4 Methods

2.4.1 Data acquisition

Subject and task

All protocols were in accord with the National Institutes of Health guidelines and approved by the Columbia University Institutional Animal Care and Use Committee. Subject C was an adult, male macaque monkey (*Macaca mulatta*) weighing 13 kg.

During experiments, the monkey sat in a primate chair with his head restrained via surgical implant and his right arm loosely restrained. To perform the task, he grasped a handle with his left hand while resting his forearm on a small platform that supported the handle. Once he had achieved a comfortable position, we applied tape around his hand and velcro around his forearm. This ensured consistent placement within and between sessions. The handle controlled a manipulandum, custom made from aluminum (80/20 Inc.) and connected to a ball bearing carriage on a guide rail (McMaster-Carr, PN 9184T52). The carriage was fastened to a load cell (FUTEK, PN FSH01673), which was locked in place. The load cell converted one-dimensional (tensile and compressive) forces to a voltage signal. That voltage was amplified (FUTEK, PN FSH03863) and routed to a Performance real-time target machine (Speedgoat) that executed a Simulink model (MathWorks) to run the task. As the load cell was locked in place, forces were applied to the manipulandum via isometric contractions.

The monkey controlled a 'Pac-Man' icon, displayed on an LCD monitor (Asus PN PG258Q, 240 Hz refresh, 1920 x 1080 pixels) using Psychophysics Toolbox 3.0. Pac-Man's horizontal position was fixed on the left hand side of the screen. Vertical position was directly proportional to the force registered by the load cell. For 0 Newtons applied force, Pac-Man was positioned at the bottom of the screen; for the calibrated maximum requested force for the session, Pac-Man was positioned at the top of the screen. Maximum requested forces (see: Experimental Procedures,

below) were titrated to be comfortable for the monkey to perform across multiple trials and to activate multiple MUs, but not so many that rendered EMG signals unsortable. On each trial, a series of dots scrolled leftwards on screen at a constant speed (1344 pixels/s). The monkey modulated Pac-Man's position to intercept the dots, for which he received juice reward. Thus, the shape of the scrolling dot path was the temporal force profile the monkey needed to apply to the handle to obtain reward. We trained the monkey to generate static, step, ramp, and sinusoidal forces over a range of amplitudes and frequencies. We define a 'condition' as a particular target force profile (e.g., a 2 Hz sinusoid) that was presented on many 'trials', each a repetition of the same profile. Each condition included a 'lead-in' and 'lead-out' period: a one-second static profile appended to the beginning and end of the target profile, which facilitated trial alignment and averaging (see below). Trials lasted 2.25-6 seconds, depending on the particular force profile. Juice was given throughout the trial so long as Pac-Man successfully intercepted the dots, with a large 'bonus' reward given at the end of the trial.

The reward schedule was designed to be encouraging; greater accuracy resulted in more frequent rewards (every few dots) and a larger bonus at the end of the trial. To prevent discouraging failures, we also tolerated small errors in the phase of the response at high frequencies. For example, if the target profile was a 3 Hz sinusoid, it was considered acceptable if the monkey generated a sinusoid of the correct amplitude and frequency but that led the target by 100 ms. To enact this tolerance, the target dots sped up or slowed down to match his phase. The magnitude of this phase correction scaled with the target frequency and was capped at ± 3 pixels/frame. To discourage inappropriate strategies (e.g., moving randomly, or holding in the middle with the goal if intercepting some dots) a trial was aborted if too many dots were missed (the criterion number was tailored for each condition).

Surgical procedures

After task performance stabilized at a high level, we performed a sterile surgery to implant a cylindrical chamber (Crist Instrument Co., 19 mm inner diameter) that provided access to M1. Guided by structural magnetic resonance imaging scans taken prior to surgery, we positioned the chamber surface-normal to the skull, centered over the central sulcus. We covered the skull within the cylinder with a thin layer of dental acrylic. Small (3.5 mm), hand-drilled burr holes through the acrylic provided the entry point for electrodes.

Intracortical recordings and microstimulation

Neural activity was recorded with Neuropixels probes. Each probe contained 128 channels (two columns of 64 sites). Probes were lowered into position with a motorized microdrive (Narishige). To facilitate inserting the probe, a 'sharp' microelectrode (FHC) was used to poke through the dura by hand before lowering the Neuropixels probe. Pre-poking the dura was done with the metal guide tube in place (resting on the surface of the cortical tissue) and the Neuropixels probe floating above and off-axis from the guide tube. Pre-poking was often sufficient to ensure smooth entry into cortex with the Neuropixels probe. Recordings were made at depths ranging from 5.6 - 12.1 mm relative to the surface of the dura. Raw neural signals were digitized at 30 kHz and saved with a 128-channel neural signal processor (Blackrock Microsystems, Cerebus).

Intracortical electrical stimulation (20 biphasic pulses, 333 Hz, 400 s phase durations, 200 s interphase) was delivered through linear arrays (Plexon Inc., S-Probes) using a neurostimulator (Blackrock Microsystems, Cerestim R96). Each probe contained 32 electrode sites with 100 m separation between them. Probes were positioned with a motorized microdrive (Narishige). We estimated the target depth by recording neural activity prior to stimulation sessions. Each stimulation experiment began with an initial mapping, used to select 4-6 electrode sites to be used in the experiments. That mapping allowed us to estimate the muscles activated from each site, and the associated thresholds. Thresholds were determined based on visual observation and were typically

low (10-50 μA), but occasionally quite high (100-150+ μA) depending on depth. Across all 32 electrodes, microstimulation induced twitches of proximal and distal muscles of the upper arm, ranging from the deltoid to the forearm. Rarely did an electrode site fail to elicit any response, but many responses involved multiple muscles or gross movements of the shoulder that were difficult to attribute to a specific muscle. Yet some sites produced more localized responses, prominent only within a single muscle head. Sometimes a narrow (few mm^2) region within the head of one muscle would reliably and visibly pulse following stimulation. Because penetration locations were guided by recordings and stimulation on previous days, such effects often involved the muscles central to performance of the task: the *deltoid* and *triceps*. In such cases, we selected 4-6 sites that produced responses in one of these muscles, and targeted that muscle with EMG recordings. EMG recordings were always targeted to a localized region of one muscle head (see below). In cases where stimulation appeared to activate only part of one muscle head, EMG recordings targeted that localized region.

EMG recordings

Intramuscular EMG activity was recorded acutely using paired hook-wire electrodes (Natus Neurology, PN 019-475400). Electrodes were inserted ~ 1 cm into the muscle belly using 30 mm x 27 G needles. Needles were promptly removed and only the wires remained in the muscle during recording. Wires were thin (50 μm diameter) and flexible and their presence in the muscle is typically not felt after insertion, allowing the task to be performed normally. Wires were removed at the end of the session.

We employed several modifications to facilitate isolation of MU spikes. As originally manufactured, two wires protruded 2 mm and 5 mm from the end of each needle (thus ending 3 mm apart) with each wire insulated up to a 2 mm exposed end. We found that spike sorting benefited from including 4 wires per needle (i.e., combining two pairs in a single needle), with each pair having a differently modified geometry. Modifying each pair differently meant that they tended

to be optimized for recording different MUs[129]; one MU might be more prominent on one pair and the other on another pair. Electrodes were thus modified as follows. The stripped ends of one pair were trimmed to 1 mm, with 1 mm of one wire and 8 mm of the second wire protruding from the needle's end. The stripped ends of the second pair were trimmed to 0.5 mm, with 3.25 mm of one wire and 5.25 mm of the second wire protruding. Electrodes were hand fabricated using a microscope (Zeiss), digital calipers, precision tweezers and knives. During experiments, EMG signals were recorded differentially from each pair of wires with the same length of stripped insulation; each insertion thus provided two active recording channels. Four insertions (closely spaced so that MUs were often recorded across many pairs) were employed, yielding eight total pairs. The above approach was used for both the dynamic and muscle-length experiments, where a challenge was that normal behavior was driven by many MUs, resulting in spikes that could overlap in time. This was less of a concern during the microstimulation experiments. Stimulation-induced responses were typically fairly sparse near threshold (a central finding of our study is that cortical stimulation can induce quite selective MU recruitment). Thus, microstimulation experiments employed one electrode pair per insertion, with minimal modification (exposed ends shorted to 1 mm).

Raw voltages were amplified and analog filtered (band-pass 10 Hz - 10 kHz) with ISO-DAM 8A modules (World Precision Instruments), then digitized at 30 kHz with a neural signal processor (Blackrock Microsystems, Cerebus). EMG signals were digitally band-pass filtered online (50 Hz - 5 kHz) and saved.

Experimental procedures

Cortical recordings were performed exclusively during one set of experiments ('dynamic', defined below), whereas EMG recordings were conducted across three sets of experiments (dynamic, 'muscle length', and microstimulation). In a given session, the eight EMG electrode pairs were inserted within a small (typically ~ 2 cm^2) region of a single muscle head. This focus aided sorting by ensuring that a given MU spike typically appeared, with different waveforms, on multiple chan-

nels. This focus also ensured that any response heterogeneity was due to differential recruitment among neighboring MUs.

In dynamic experiments, the monkey generated a diverse set of target force profiles. The manipulandum was positioned so that the angle of shoulder flexion was 25° and the angle of elbow flexion was 90°. Maximal requested force was 16 Newtons. We employed twelve conditions (**fig. S2.6**) presented interleaved in pseudo-random order: a random order was chosen, all conditions were performed, then a new random order was chosen. Three conditions employed static target forces: 33%, 66% and 100% of maximal force. Four conditions employed ramps: increasing or decreasing across the full force range, either fast (lasting 250 ms) or slow (lasting 4 s). Four conditions involved sinusoids at 0.25, 1, 2, and 3 Hz. The final condition was a 0-3 Hz chirp. The amplitude of all sinusoidal and chirp forces was 75% of maximal force, except for the 0.25 Hz sinusoid, which was 100% of maximal force. Recordings in dynamic experiments were made from the deltoid (typically the anterior head and some from the lateral head) and the triceps (lateral head).

In muscle-length experiments, the monkey generated force profiles with his deltoid at a long or short length (relative to the neural position used in the dynamic experiments). The manipulandum was positioned so that the angle of shoulder flexion was 15° (long) or 50° (short), while maintaining an angle of elbow flexion of 90°. Maximal requested forces were 18 N (long) and 14 N (short). Different maximal forces were employed as it appeared more effortful to generate forces in the shortened position. To ensure enough trials per condition, we employed only a subset of the force profiles used in the dynamics experiments. These were 2 static forces (50% and 100% of maximal force), the slow increasing ramp, both increasing and decreasing fast ramps, all four sinusoids and the chirp. These were presented interleaved in pseudorandom order for multiple trials (~ 30 per condition) for the lengthened position (15°) before changing to the shortened position (50°). In most experiments we were able to revert to the lengthened position (15°) at the end of the session,

and verify that MU recruitment returned to the originally observed pattern. Recordings in muscle-length experiments were made from the deltoid (anterior head).

Microstimulation experiments employed recordings from the lateral deltoid and lateral triceps. Both these muscles exhibited strong task-modulated activity, as documented in the dynamic and muscle-length experiments. We also included recordings from the sternal pectoralis major, which showed only modest task-modulated activity, as we found cortical sites that reliably activated it. The manipulandum was positioned so that the angle of shoulder flexion was 25° and the angle of elbow flexion was 90° (as in dynamic experiments). Maximal force was typically set to 16 N, but was increased to 24 N and 28 N for two sessions each in an effort to evoke greater muscle activation.

Microstimulation experiments employed a limited set of force profiles: four static forces (0, 25%, 50% and 100%), and the slow (4 s) increasing ramp. The ramp was included to document the natural recruitment pattern during slowly changing forces. Microstimulation was delivered once per trial during the static forces, at a randomized time (1000-1500 ms relative to when the first dot reached Pac-Man). Because stimulation evoked activity in muscles used to perform the task, it sometimes caused small but detectable changes in force applied to the handle. However, these were so small that they did not impact the monkey's ability to perform the task and appeared to go largely unnoticed. These experiments involved a total of 17-25 conditions: the ramp condition (with no stimulation) plus the four static forces for the 4-6 chosen electrode sites. These were presented interleaved in pseudorandom order.

2.4.2 Data processing

Signal processing and spike sorting

Cortical voltage signals were spike sorted using KiloSort 2.0[130]. A total of 881 neurons were isolated across 15 sessions.

EMG signals were digitally filtered offline using a second-order 500 Hz high-pass Butterworth. Any low SNR or dead EMG channels were omitted from analyses. Motor unit (MU) spike times were extracted using a custom semi-automated algorithm. As with standard spike-sorting algorithms used for neural data, individual MU spikes were identified based on their match to a template: a canonical time-varying voltage across all simultaneously recorded channels (example templates are shown in **fig. 2.1d**, *bottom left*). A distinctive feature of intramuscular records (compared to neural recordings) is that they have very high signal-to-noise (peak-to-peak voltages on the order of mV, rather than μV, and there is negligible thermal noise) but it is common for more than one MU to spike simultaneously, yielding a superposition of waveforms. This is relatively rare at low forces but can become common as forces increase. Our algorithm was thus tailored to detect not only voltages that corresponded to single MU spikes, but also those that resulted from the superposition of multiple spikes. An example of this is illustrated in **fig. 2.1d** (*bottom right*): the third spike from MU12 coincides with that of MU1, and both are successfully detected. Detection of superposition was greatly aided by the multi-channel recordings; different units were prominent on different channels. Further details are provided in **section 2.6.1**.

Trial alignment and averaging

Single-trial spike rasters, for a given neuron or MU, were converted into a firing rate via convolution with a 25 ms Gaussian kernel. One analysis (**fig. 2.4d**) focused on single-trial responses, but most employed trial-averaging to identify a reliable average firing rate. To do so, trials for a given condition were aligned temporally and the average firing rate, at each time, was computed across trials. Stimulation trials were simply aligned to stimulation onset. For all other conditions, each trial was aligned on the moment the target force profile 'began' (when the target force profile, specified by the dots, reached Pac-Man). This alignment brought the actual (generated) force profile closely into register across trials. However, because the actual force profile could sometimes slightly lead or lag the target force profile, some modest across-trial variability remained. Thus, for all trials with changing forces, we realigned each trial (by shifting it slightly in time) to min-

imize the mean squared error between the actual force and the target force profile. This ensured that trials were well-aligned in terms of the actual generated forces (the most relevant quantity for analyses of MU activity). Trials were excluded from analysis if they could not be well aligned despite searching over shifts from -200 to 200 ms.

2.4.3 Data Analysis

Quantifying motor unit flexibility

We developed two analyses that quantified MU-recruitment flexibility without directly fitting a model (model-based quantification is described below). These two analyses were used to produce the results in **figs. S2.8** and **S2.9**, respectively. Both methods leverage the definition of rigid control to detect patterns of activity that are inconsistent with rigid control even under the most generous of assumptions.

Let $\mathbf{r}_t = \begin{bmatrix} r_{1,t} & r_{2,t} & \cdots & r_{n,t} \end{bmatrix}^\top$ denote the population state at time t, where $r_{i,t}$ denotes the firing rate of the i^{th} MU. If $\mathbf{r}_t$ traverses a 1-D monotonic manifold, then as the firing rate of one MU increases, the firing rate of all others should either increase or remain the same. More generally, the change in firing rates from t to t' should either be nonnegative or nonpositive for all MUs. If the changes in firing rate were all nonnegative with some increases, then we could infer that a common input drive increased from t to t'. Equivalently, we could conclude that the common drive decreased from t' to t. Both these cases (all nonnegative or all nonpositive) are consistent with rigid control because there exists some 1-D monotonic manifold that contains the data at both t' and t.

On the other hand, departures from a 1-D monotonic manifold can be inferred as moments when the firing rates of one or more MUs increase as others' decrease. Both our analyses seek to quantify the magnitude of such departures while being very conservative. Specifically, the size of a departure was always measured as the smallest possible discrepancy from a 1-D manifold, based

on all possible 1-D manifolds. To illustrate the importance of this conservative approach, consider a situation where the firing rate of MU1 increases considerably while MU2's rate decreases slightly from t to t'. This scenario would be inconsistent with activity being modulated solely by a common input, yet it would be impossible to know which MU reflected an additional or separate input. Perhaps common drive decreased slightly (explaining the slight decrease in MU2's rate) but MU1 received an additional large, private excitatory/inhibitory input. This would indicate a large departure from rigid control. Yet another possibility is that common drive increased considerably (explaining the large increase in MU1's rate) and that MU2's rate failed to rise because it was already near maximal firing rate. This would not explain why MU2's rate went down, but if that decrease was small it could conceivably be due to a very modest departure from idealized rigid control. Thus, to be conservative, one should quantify this situation as only a slight deviation from the predictions of rigid control. Both methods described below were designed to do so; when MU activities were anticorrelated, we identified the largest increase and decrease in firing rates, then reported the change that was smaller in magnitude.

For the first analysis, we computed the largest nonnegative change in firing rates from t to t' for a population of n MUs as

$$\Delta r^{+}(t, t') = \max\left(0, r_{1,t} - r_{1,t'}, r_{2,t} - r_{2,t'}, \ldots, r_{n,t} - r_{n,t'}\right). \tag{2.1}$$

If a 1-D monotonic manifold can be drawn through $\mathbf{r}_t$ and $\mathbf{r}_{t'}$, then either $\Delta r^{+}(t, t')$ or $\Delta r^{+}(t', t)$ will be zero. Otherwise, $\Delta r^{+}(t, t')$ will capture the largest increase (across MUs) in rate from t to t' while $\Delta r^{+}(t', t)$ will capture the largest decrease. Thus, we computed departures from a monotonic manifold at the level of an individual MU as

$$D(t, t') = \min\left(\Delta r^{+}(t, t'), \Delta r^{+}(t', t)\right). \tag{2.2}$$

As examples, consider a population of two MUs with $\mathbf{r}_t = [10, 10]$ and $\mathbf{r}_{t'} = [15, 25]$. These

states would be consistent with an increase in common drive from t to t', so $D(t, t') = 0$ (**fig. S2.8a**, *left*). Conversely, $\mathbf{r}_t = [10, 10]$ and $\mathbf{r}_{t'} = [9, 30]$ (**fig. S2.8a**, *center*) suggests a violation of rigid control, but that violation might be small; one can draw a manifold that passes through $[10, 10]$ and comes within 1 spike/s of $[9, 30]$. In this case, $D(t, t') = 1$. Finally, $\mathbf{r}_t = [10, 10]$ and $\mathbf{r}_{t'} = [0, 30]$ (**fig. S2.8a**, *right*) argue for a sizable violation; $[0, 30]$ is at least 10 spikes/s distant from any monotonic manifold passing through $[10, 10]$, so $D(t, t') = 10$.

It is worth emphasizing that **eq.** (2.2) can readily be computed for a population with more than two MUs, but the analysis ultimately reduces to a comparison of two MUs: one whose firing rate increased the most and the other whose firing rate decreased the most across a pair of time points.

To extend our analysis to multiple time points, we computed the 'MU displacement' as

$$d_{\mathrm{MU}}(t) = \min_{\tau,\tau'} \left(\max_{t'} D(t + \tau, t' + \tau') \right) \tag{2.3}$$

where t' indexes over all other times and conditions, and τ and τ' are time lags. The inclusion of time lags ensures that departures from a monotonic manifold cannot simply be attributed to modest differences in response latencies across MUs. In our analyses, we optimized over $\tau, \tau' \in [-25, 25]$ ms. d_{MU} is exceedingly conservative; it makes no assumptions regarding the manifold other than that it is monotonic, and identifies only those violations that are apparent when comparing just two times.

An advantage of the d_{MU} metric is interpretational simplicity; it identifies pairs of times where the joint activity of two MUs cannot lie on a single 1-D monotonic manifold. A disadvantage is that it does not also capture the degree to which multiple other MUs might also have activity inconsistent with a 1-D monotonic manifold. To do so, we employed a second metric that quantifies MU-recruitment flexibility at the population level. Under the assumptions of rigid control, the

magnitude of common drive determines the population state and therefore the summed activity of all MUs or, equivalently, its L1-norm, $\|\mathbf{r}\|_1$. Increases and decreases in common drive correspond, in a one-to-one manner, to increases and decreases in $\|\mathbf{r}\|_1$.Violations of rigid control can thus be inferred if a particular norm value, λ, is associated with different population states. Geometrically, this corresponds to the population activity manifold intersecting the hyperplane defined by $\|\mathbf{r}\|_1 = \lambda$ at multiple locations.

We thus defined the motor neuron pool (MNP) dispersion as

$$d_{\mathrm{MNP}}(\lambda) = \min_{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2} \left(\max_{t_1,t_2 \in \Omega} \|\mathbf{r}_{t_1+\tau_1} - \mathbf{r}_{t_2+\tau_2}\|_1 \right), \quad \Omega = \{t \,:\, |\|\mathbf{r}_t\|_1 - \lambda| < \varepsilon\} \tag{2.4}$$

where $\boldsymbol{\tau}, \boldsymbol{\tau}_2$ are time lag vectors, of the same dimensionality as $\mathbf{r}$, and ε is a small constant. Conceptually, the dispersion identifies the pair of time points when the population states are the most dissimilar, while having norms within ε of λ. As when computing $d_{\mathrm{MU}}(t)$, we minimized $d_{\mathrm{MNP}}(\lambda)$ over time lags so as to only consider dispersions that could not be simply attributed to latency differences across MUs. For our analyses, we set $\varepsilon = 1$ and optimized over $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \in [-25, 25]$ ms.

Latent factor model

We developed a probabilistic latent variable model of MU activity. Let x_t be the unknown latent variables at time t, which are shared between all MUs. We can fit this model with one latent (**fig. 2.4**; x_t can be a single value) or multiple latents (**fig. 2.5**). Let $y_{i,t}$ be the activity of the i^{th} MU at time t, given by

$$y_{i,t} \sim \mathcal{N}(f_i(x_{t+\tau_i}), \epsilon) \tag{2.5}$$

where f_i denotes the link function for the i^{th} MU and τ_i denotes the lag between its response and the shared latent variables. We constrained $\tau_i \in [-25, 25]$ ms. To identify flexible, monotonically increasing link functions with nonnegative outputs, we parameterized f_i as a rectified monotonic neural network. More precisely, we fit each f_i using a two-layer feedforward neural network in

which the weights were constrained to be positive. The positivity constraint was achieved by letting each weight $w = \ln(1 + e^u)$, where the values of u were fit within the model. During model training, the output of the neural network was passed through a 'leaky rectified linear unit' (i.e., so that the output was never exactly zero). After training was completed, we used standard rectification on the output.

When predicting held-out data, we encouraged temporal smoothness in the latent space to improve generalization performance by letting $\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_{t-1}, \sigma)$, where smaller values of σ encouraged greater smoothness. We set σ to 0.01 for our analyses.

To infer the most likely distribution of latent variables given the data (i.e., the model posterior, $p(\mathbf{x}|\mathbf{y})$), and to learn the link functions and other parameters, we used variational inference with a mean-field approximation for the posterior approximation. As an inference method, we used black-box variational inference[122], which performs gradient descent to maximize the model's evidence lower bound. We iterated between (1) optimizing the posterior and all parameters while holding response lags fixed and (2) optimizing the response lags. Post model-fitting, when predicting MU activity, we used the mean of the posterior distribution as the latent input at each time.

Prior to fitting the model, the firing rate of each MU was normalized by its maximum response across conditions. Normalization did not alter the ability of the model to fit the data, but simply encouraged the model to fit all MUs, rather than just the high-rate units. Additionally, the likelihood of each time point was weighted by the duration of the experimental condition, so that each condition mattered equally within the model regardless of duration. When fitting to single trials, we also weighted each condition by its trial count, again so that each condition had equal importance. All model fits were done within individual sessions.

Residual error plots

To compute the cross-validated model residuals, we first randomly split the single-trial firing rates for each MU into halves, and computed the trial-average responses for each half: $\mathbf{y}_{i,1}$ and $\mathbf{y}_{i,2}$. We then fit the latent variable model to each half, which yielded a pair of predicted responses, $\hat{\mathbf{y}}_{i,1}$ and $\hat{\mathbf{y}}_{i,2}$. The cross-validated model residuals were calculated as the dot product between the residual errors of each half: $(\mathbf{y}_{i,1} - \hat{\mathbf{y}}_{i,1})^\top(\mathbf{y}_{i,2} - \hat{\mathbf{y}}_{i,2})$. We computed the median cross-validated residuals across all MUs and sessions for a given partitioning of the data. The above steps were then repeated for 10 different random splits of trials and we reported the mean ± standard error of the median error across re-partitions and fits.

As a control (**fig. 2.4c**), we modified the data so that a single latent variable could fully account for all responses. To do so, we reconstructed the firing rates using only the first principle component of the trial average firing rates. For example, if $\mathbf{w}$ is the $n \times 1$ loading vector for the first principal component, then Y_1, the $ct \times n$ matrix of responses for one partitioning of the data, was reconstructed as $[Y_1\mathbf{w}\mathbf{w}^\top]_+$, where the rectification ensures that all firing rates are non-negative. Using these reconstructed firing rates, we performed the same residual error analysis. Because of the rectification, the modified data are not one-dimensional in the linear sense (there would be multiple principal components with non-zero variance). Yet because the data will lie on a one-dimensional monotonic manifold, cross-validated error should be near zero when fitting the model, which is indeed what we observed.

Consistency plots

We fit the model to the activity of single trials. We aimed to determine whether, when fit to two conditions, the model consistently overestimated the true firing rates in one condition and underestimated the firing rates in the other condition. To do so, we calculated the mean model error across time on every trial for each condition. Let $E(1, tr)$ and $E(2, tr)$ denote the mean errors for a particular MU, pair of conditions (indexed by 1 and 2), and trial tr. We calculated the consistency

for the MU and conditions as

$$C = \max\left[\left(\frac{n_{\text{over, 1}} + n_{\text{under, 2}} + 0.5 \cdot n_{\text{equal}}}{n}\right), \left(\frac{n_{\text{under, 1}} + n_{\text{over, 2}} + 0.5 \cdot n_{\text{equal}}}{n}\right)\right] \tag{2.6}$$

where

$$n_{\text{over, } j} = \sum_{tr} \mathbf{1}_{E(j,tr)>0}$$
$$n_{\text{under, } j} = \sum_{tr} \mathbf{1}_{E(j,tr)<0}$$
$$n_{\text{equal}} = \sum_{tr} \mathbf{1}_{E(1,tr)=0} + \sum_{tr} \mathbf{1}_{E(2,tr)=0}$$

n is the total number of trials across both conditions, and $\mathbf{1}_A$ is the indicator function (1 if A is true; 0 otherwise). **Eq.** (2.6) determines the fraction of times one condition had negative errors and the other had positive errors, while accounting for trials with no error. Prior to performing this consistency calculation, we set all $E(j, tr)$ with absolute value less than 0.01 to 0, so that the sign of negligible errors was not considered. We also removed $E(1, tr)$ or $E(2, tr)$ in which the MU had zero actual and predicted activity, because it was impossible for the predicted activity to undershoot the true activity in this setting.

We calculated the fraction of MUs that had $C > 0.8$ and an average error of at least 0.01 across trials (to ensure that outlier trials did not lead to false positives of consistent errors). We excluded MUs who had zero activity in > 80% of trials in the two conditions being analyzed. Consequently, the number of MUs included in the analysis) varied for each pair of conditions.

To calculate a chance-level baseline (**fig. 2.4d**), for each MU, we calculated the probability that greater than 80% of the included trials would have a positive or negative error, assuming that each trial has an independent 50/50 chance of being positive or negative. More precisely, let $F(k; n, p)$ be the cumulative density function of a binomial distribution of having k successes in

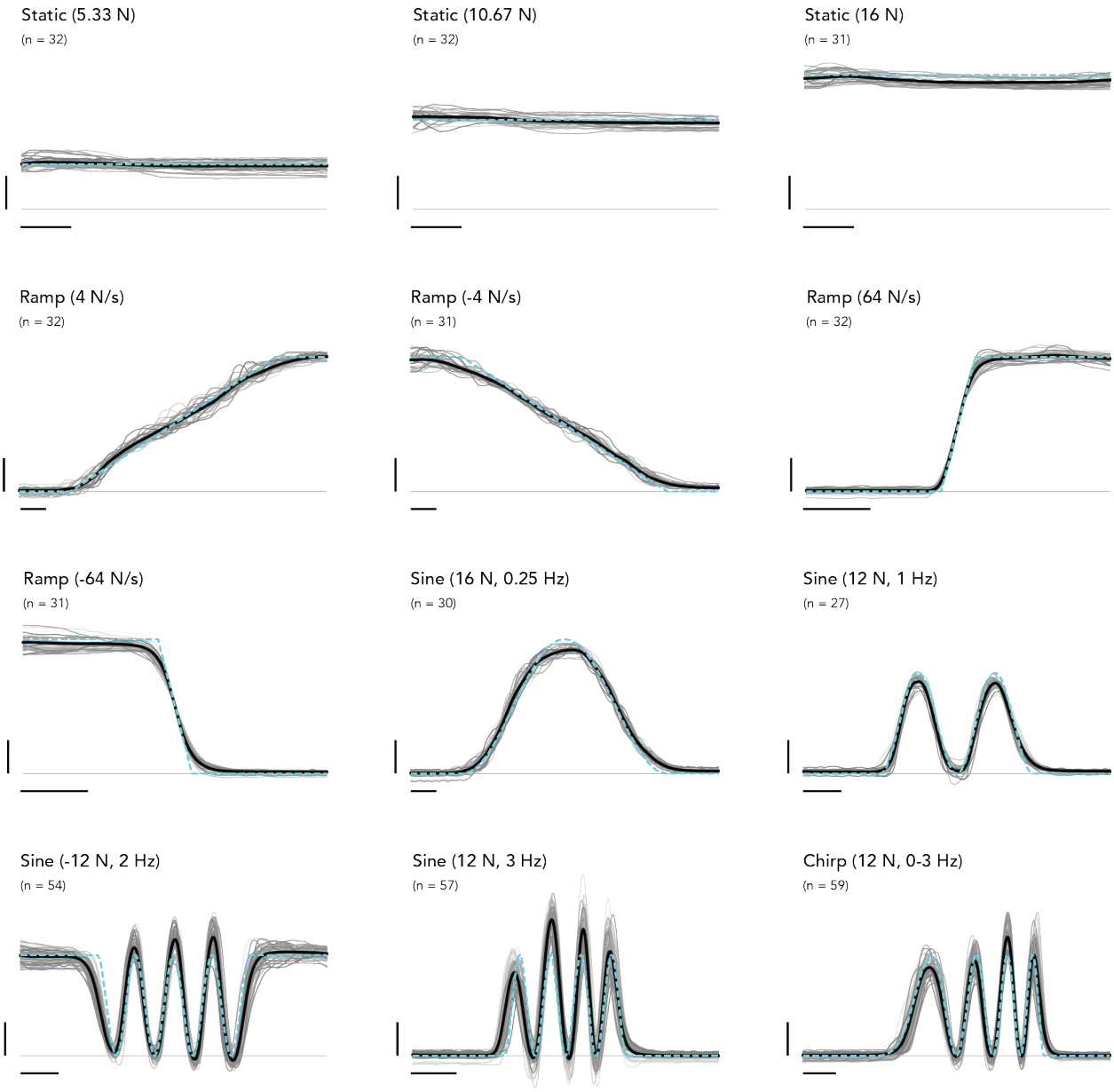
n Bernoulli events, each event with probability p of being a success. We calculate $P_i = 2(1 - F(\text{ceil}[0.8n_i]; n_i, 0.5))$, where n_i is the number of total trials included for MU i and ceil[] gets the next integer. The total expected fraction of MUs with $C > 0.8$ by chance is thus $\sum_i P_i$.

Cross-validated reliability dimensionality estimate

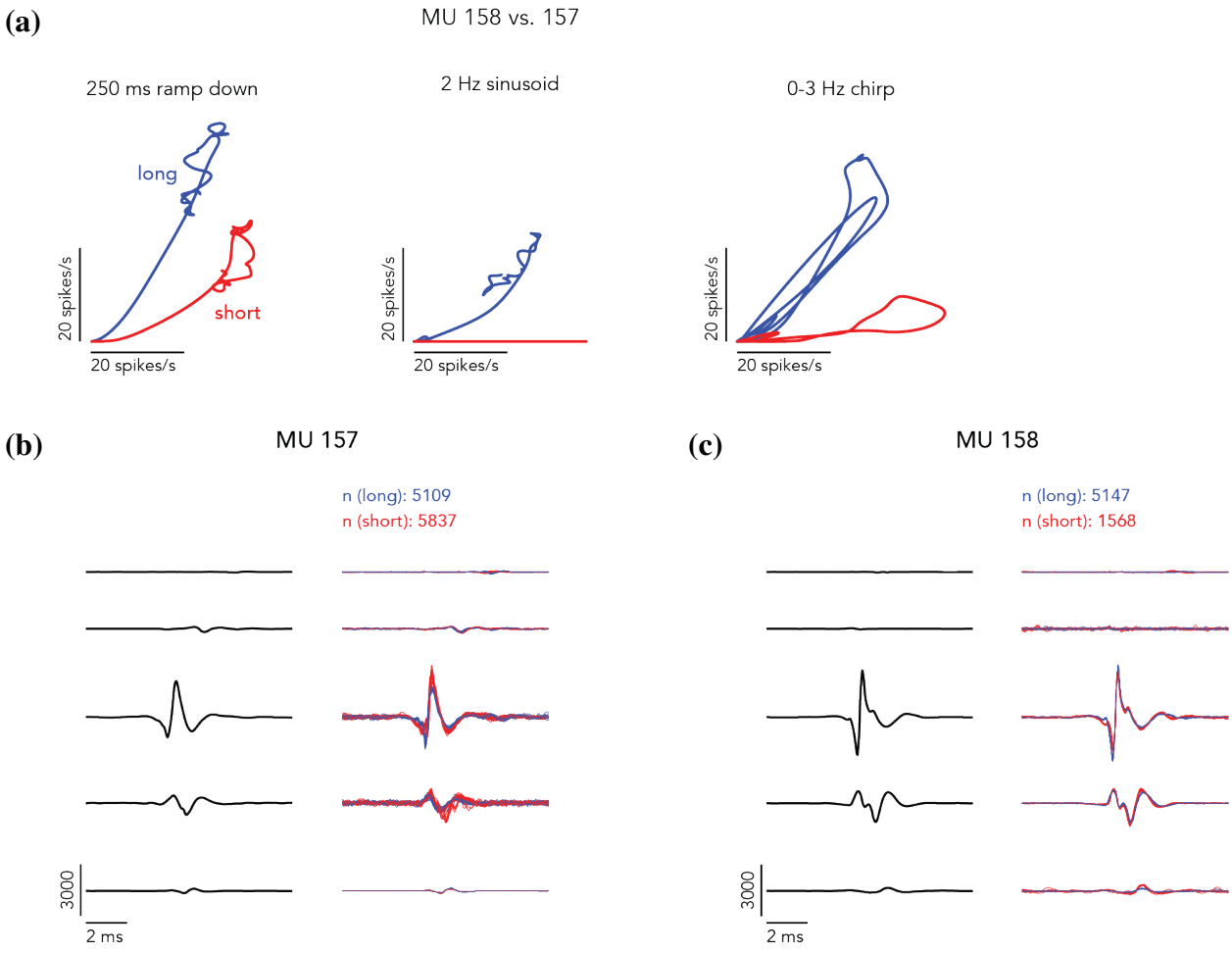
To estimate the dimensionality of M1, we randomly split the single-trial firing rates for each neuron into two groups and averaged over trials within each group. Let Y_1 and Y_2 denote the $CT \times N$ matrices of trial-averaged responses for each partition (CT condition-times and N neurons). Let $\mathbf{w}_i$ (an $N \times 1$ vector) denote the i^{th} principal component (PC) of Y_1. The reliability of PC i was computed as the correlation between $Y_1\mathbf{w}_i$ and $Y_2\mathbf{w}_i$. We repeated this process for 25 re-partitions over trials to obtain confidence intervals. Our method is inspired by Churchland *et al.*[22] and conceptually similar to but distinct from the cross-validated PCA analysis of Stringer *et al.*, which estimates the stimulus-related ('signal') neural variance based on spontaneous activity across many neurons on single trials[124].

To create simulated data sets with dimensionality k, we computed $Y^k = YQ_kQ_k^\top$, where Y is the matrix of M1 firing rates averaged over all trials, and Q_k denotes the first k columns of a random orthonormal matrix. Simulated single-trial spikes were generated for each neuron using an inhomogeneous Poisson process with rate given by the corresponding column of Y^k. Simulated spikes were smoothed using a 25 ms Gaussian kernel, and the cross-validated reliability metric applied as described above.

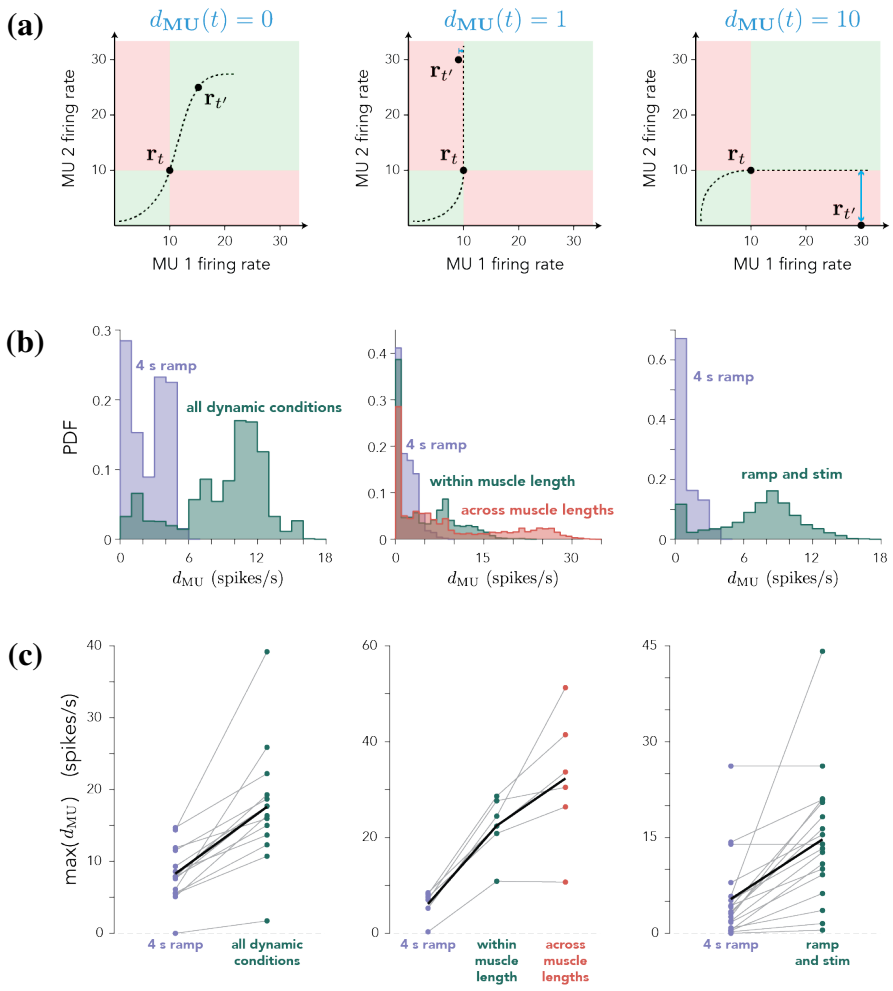
2.5 Supplementary Figures



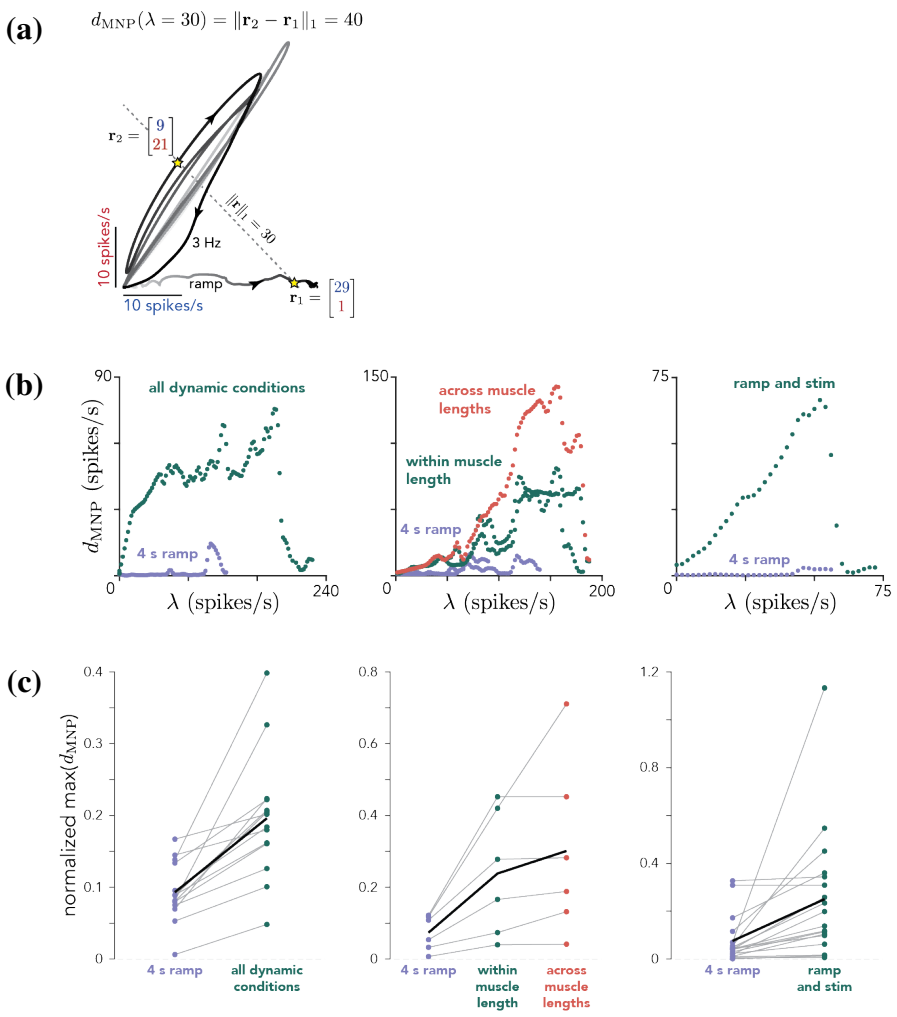
Supplementary Figure S2.6: Force profiles. Single-trial (*gray*), trial-averaged (*black*), and target (*cyan*) forces for one session of dynamic experiments. Vertical scale bars indicate 4 N. Horizontal scale bars indicate 500 ms. *n* denotes trial count.



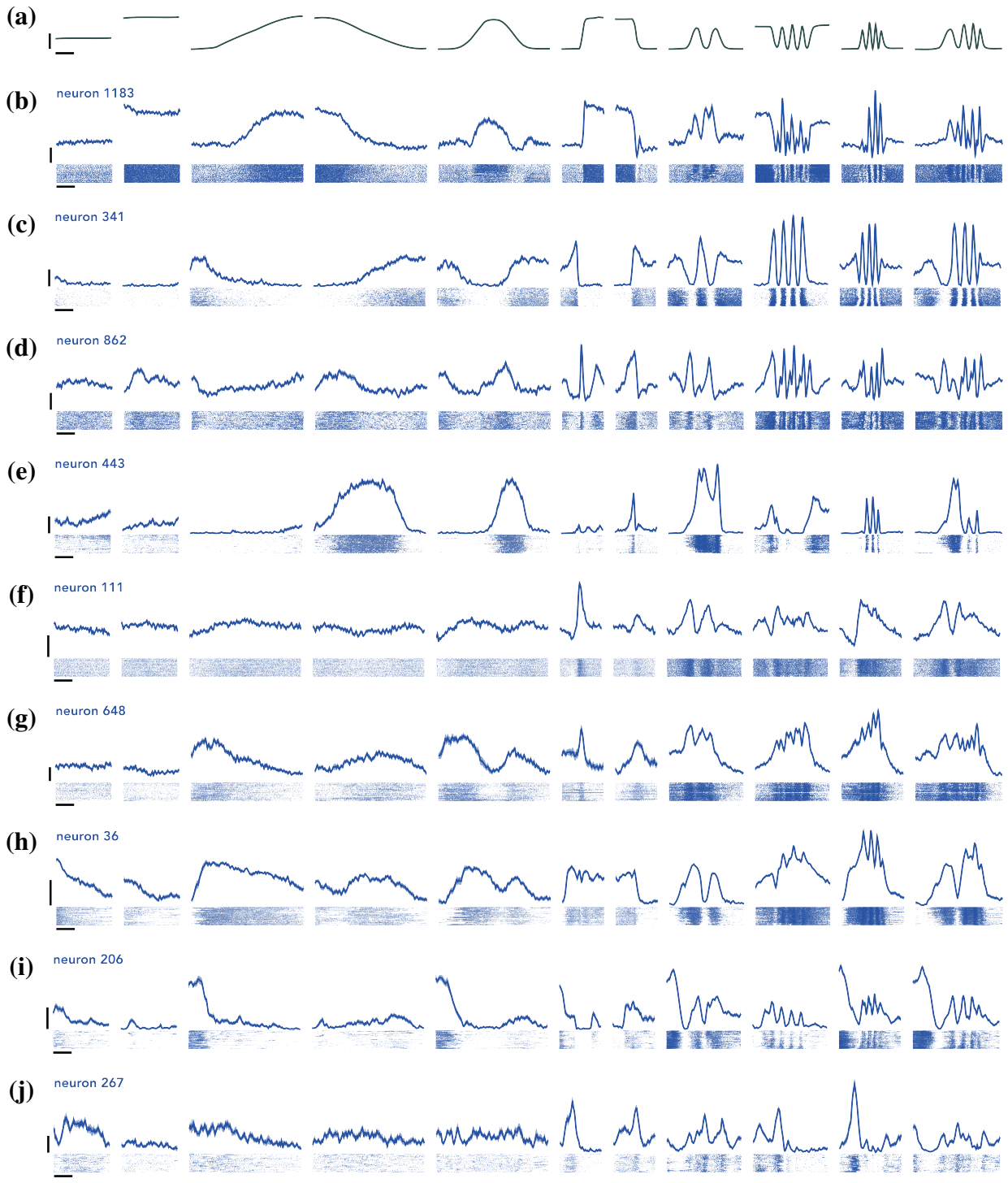
Supplementary Figure S2.7: Example MU responses and waveforms across muscle lengths. (**a**) Firing rate of a pair of simultaneously recorded deltoid MUs plotted against each other for three different conditions (*columns*) with the deltoid in a lengthened (*blue*) or shorted (*red*) posture. (**b**) *Left*. Template of MU157 across the 5 EMG channels used during this session. *Right*. The 20 waveforms identified in each posture that were most similar to the template. *n* denotes the total spike counts in each posture. (**c**) Same as ***b*** for MU158.



Supplementary Figure S2.8: MU displacement. (**a**) Schematic illustrating, in three situations, the size of the displacement (d_{MU}) for a two-dimensional population state at two times. *Left.* $d_{\mathrm{MU}}(t) = 0$ because a monotonic manifold can pass through $\mathbf{r}_t$ and $\mathbf{r}_{t'}$. *Center.* Any monotonic manifold passing through $\mathbf{r}_t$ is restricted to the green zone, and thus cannot come closer than 1 spike/s to $\mathbf{r}_{t'}$. *Right.* A manifold passing through $\mathbf{r}_t$ can come no closer than 10 spikes/s from $\mathbf{r}_{t'}$. (**b**) Probability density function (PDF) of $d_{\mathrm{MU}}(t)$ for one session for each experiment. $d_{\mathrm{MU}}(t)$ was evaluated at every time during the 4 s increasing ramp condition alone, at one muscle length, (*purple*) or including other conditions. *Left.* For dynamic experiments, the other conditions were the different force profiles. *Center.* For muscle-length experiments, the other conditions were different force profiles using the same muscle length (a subset of the force profiles used in the dynamic experiments) or all force profiles across both muscle lengths. *Right.* For microstimulation experiments, the other conditions involved cortical stimulation (on one of 4-6 electrodes) during static force production at different levels. (**c**) Maximum displacement (across time) for each condition group shown in ***b*** for all sessions. Thin gray lines correspond to different sessions and the thick black line corresponds to the mean across sessions.



Supplementary Figure S2.9: Motor neuron pool (MNP) dispersion. (**a**) Firing rate of MU309 vs. MU311 during the 4 s increasing ramp and 3 Hz sinusoidal conditions. The line defined by $\|\mathbf{r}\|_1 = 30$ intercepts the activity manifold at several different moments; of those, $\mathbf{r}_1$ and $\mathbf{r}_2$ are the most separated along the contour line. The MNP dispersion for $\lambda = 30$ is the L1-norm of the difference between $\mathbf{r}_1$ and $\mathbf{r}_2$: 40 spikes/s. (**b**) Scatter plot of d_{MNP} versus λ for one session for each experiment. d_{MNP} was evaluated at every time during the 4 s increasing ramp condition alone, at one muscle length, (*purple*) or including other conditions. *Left*. For dynamic experiments, the other conditions were the different force profiles. *Center*. For muscle-length experiments, the other conditions were different force profiles using the same muscle length (a subset of the force profiles used in the dynamic experiments) or all force profiles across both muscle lengths. *Right*. For microstimulation experiments, the other conditions involved cortical stimulation (on one of 4-6 electrodes) during static force production at different levels. (**c**) Maximum dispersion (across λ) for each condition group shown in b for all sessions. For each session, $d_{MNP}(\lambda)$ was restricted to the greatest common λ across all condition sets before computing the maximum. Maximum dispersions were normalized by the maximum L1-norm of the MNP response across all times/conditions. Thin gray lines correspond to different sessions and the thick black line corresponds to the mean across sessions.



Supplementary Figure S2.10: Example primary motor cortex (M1) neuron responses. (**a**) Trial-averaged forces for 11 of 12 conditions (intermediate static force condition is omitted for space). Vertical scale bar indicate 8 N. Horizontal scale bar indicates 1 s. (**b**-**j**) Trial-averaged firing rates with standard error (*top*) and single-trial spike rasters (*botom*). Vertical scale bars indicate 20 spikes/s. Horizontal scale bars indicate 1 s.

2.6 Supplementary Materials

2.6.1 EMG Signal Decomposition

Introduction

An impulse emitted by an α-motoneuron initiates an action potential in each of its innervated muscle fibers that propagates bidirectionally towards the tendons, driving the fibers to concurrently contract[129]. The propagating muscle fiber action potentials can be detected by an electromyographic (EMG) electrode inserted percutaneously in the belly of the muscle. The EMG signal registers the fiber potentials as a brief waveform whose shape depends on multiple factors (including the number and physiological characteristics of the fibers, their distance from the electrode and their geometric arrangements), but typically remains constant across repeated discharges[60,131]. Synaptic transmission at the neuromuscular junction rarely fails, even during maximal muscle contractions[40], meaning that each unique waveform in the EMG signal reliably identifies the discharge of one MU.

MU spike events can be extracted from EMG signals through the process of 'spike sorting'[60]. An EMG signal can be modeled as[132]

$$x_i(t) = \sum_{\ell=0}^{L-1} \sum_{j=1}^{N} w_{i,j}(\ell) s_j(t-\ell) + \eta_i(t) \tag{2.7}$$

where $x_i(t)$ is the voltage signal recorded at (discrete) time t by the i^{th} EMG electrode, $w_{i,j}$ is the waveform of the j^{th} MU as registered on the i^{th} electrode over L time steps, $s_j(t)$ is a binary indicator of whether or not the MU emitted a spike, and $\eta_i(t)$ is additive background noise. For multi-channel EMG signals, the discharge of each MU is still captured by a single delta function train, but its single-channel waveform $w_{i,j}$ is instead modeled as a spatiotemporal filter, describing the characteristic signature of its spikes on each channel[132]. The goal of spike sorting amounts to inferring the set of spike trains, $\{_{1,2}, \ldots, _N\}$, for the N MUs detected by the EMG signal(s).

While spike sorting is not unique to EMG data, EMG signals pose some unique challenges compared to other forms of data. Spike sorting is also often applied to extracellular voltage signals recorded in the brain to extract the spike times of individual neurons. Considerable efforts have advanced spike sorting methodologies to keep up with big data[133], yielding several end-to-end algorithms that perform well for high-density electrodes implanted in cerebral tissue[119,134,135]. Yet while the principles of spike sorting data recorded from cerebral or skeletal muscle tissue are the same, the practical challenges posed by either endeavor differ substantially. Relative to MUs, action potential waveforms recorded from cerebral neurons are typically shorter in duration (1-2 ms), smaller in amplitude relative to the background noise, simpler (biphasic) and more similar in shape across neurons[136]. Moreover, for certain brain regions, neural spikes are sufficiently sparse events that the superposition of multiple action potential waveforms does not considerably hinder identifying a particular neuron, especially in the high-dimensional channel space afforded by high-density electrodes[134]. In contrast, MU spike waveforms can be quite long in duration (5-10 ms), vary in amplitude over an order of magnitude, and are often complex and multiphasic (due to variations in fiber conduction delays)[60]. Furthermore, overlapping waveforms poses the greatest challenge for spike sorting EMG signals, particularly during forceful muscle contractions. The problem is not only that the rate of coincident spiking increases with contraction intensity, due to MU recruitment; it is also that high-threshold MUs (by definition) *only* ever appear in the EMG signal superposed with other MUs. Consequently, it can be nearly impossible to ever obtain an isolated view of such MUs' characteristic waveform and thereby infer their discharge times. To extend the cocktail party problem used to analogize spike sorting[137], it can be ambiguous whether the room grew louder (EMG signal power increased) only because everyone began talking louder (MU firing rates increased) or because new, raucous people entered the room (high-threshold MUs were recruited).

Recent work using convolutive blind source separation[132] has pushed the limits of EMG de-

composition accuracy during forceful muscle contractions. However, this approach has only been tested during steady, prolonged voluntary contractions (5-30 s) and is recommended for signals lasting at least 10 seconds. In our experience, blind source separation provided accurate decomposition when forces changed slowly, but struggled to identify MUs that spiked only occasionally when forces were modulated rapidly. Thus, to meet the particular demands of the behavioral task employed in this report, we developed a novel approach to spike sorting EMG signals. Our approach is multi-step and modular, drawing from aspects of spike sorting approaches for cerebral voltage signals[117–119] with several modifications and innovations tailored to EMG signals. Our procedure fundamentally consists of two parts: obtain the best possible estimate of each MU's waveform template, then deconvolve the EMG signals with the templates to infer each unit's spike train. Below, we provide a description of each step in our pipeline ("myosort").

Myosort Pipeline

Preprocessing Raw signals were filtered with a second-order, 500 Hz high-pass Butterworth filter.

Detection Spike sorting typically begins with the detection of all putative spike events. This is often accomplished by identifying the samples on each electrode when the voltage signal exceeded some threshold[117,119]. We found that thresholding alone performed poorly for EMG signals due to the duration, complexity, and range in amplitude of MU action potential waveforms. Setting the threshold too small caused the extracted waveforms to be wildly misaligned; using too large of a threshold inevitably caused small-amplitude waveforms to go undetected that were discernible by eye. Instead, we denoised each channel, then used a peak finding algorithm to detect spike events.

Each channel was rectified, normalized, and denoised as

$$\tilde{x}_i(t) = \begin{cases} |x_i(t)| / \sigma_i & \text{if } |x_i(t)| / \sigma_i > \Theta \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, 2, \ldots, c \tag{2.8}$$

where $x_i(t)$ denotes the EMG signal recorded on the i^{th} channel at time step t and Θ is a threshold. $\sigma_i = \text{median}\,(|\mathbf{x}_i|/0.6745)$ is an estimate of the standard deviation of the background noise[117], where $\mathbf{x}_i$ denotes the EMG signal on the i^{th} channel for all time steps. Setting Θ to 6 worked for most data sets, but was dialed up or down for recordings with exceptionally low SNR or large range in waveform amplitudes. Next, we averaged $\tilde{\mathbf{x}}_i$ across channels, then smoothed the resulting vector with a Gaussian kernel with a standard deviation of 500 μs. This yielded a new vector, $\mathbf{y}$, whose peaks captured transients on timescales longer than 500 μs occurring on any $\mathbf{x}_i$. Spike indices were inferred as peak locations in $\mathbf{y}$ with peak amplitudes greater than one (`findpeaks`, MATLAB).

Alignment Let $S = \{s_1, s_2, \ldots, s_N\}$ denote the (sorted) set of spike times returned by the detection step. For each $s_k \in S$, we extracted a corresponding waveform matrix

$$W_k = \begin{bmatrix} \mathbf{w}_{1,k} \\ \mathbf{w}_{2,k} \\ \vdots \\ \mathbf{w}_{c,k} \end{bmatrix} = \begin{bmatrix} x_1(s_k - L/2) & x_1(s_k - L/2 + 1) & \cdots & x_1(s_k + L/2 - 1) \\ x_2(s_k - L/2) & x_2(s_k - L/2 + 1) & \cdots & x_2(s_k + L/2 - 1) \\ & & \vdots & \\ x_c(s_k - L/2) & x_c(s_k - L/2 + 1) & \cdots & x_c(s_k + L/2 - 1) \end{bmatrix}$$

where L denotes the waveform duration, which was set to 3 ms. Unless otherwise noted, we will use "waveform" to refer to the multi-channel waveform matrix W_k (as opposed to the single-channel waveform $\mathbf{w}_{i,k}$).

To assign W_k to a particular MU (i.e., an observation of its characteristic waveform template), it is necessary to perform some type of cluster analysis. The success of clustering largely depends on how well aligned the W_k are to one another. Aligning neural spike events is often either not needed[119] or involves shifting s_k so that it coincides with the maximum of W_k[117]. Aligning to a maximum also performed poorly for our purposes due to the multiphasic nature of MU waveforms. Instead, we developed a procedure to optimally align waveforms to one another based on their full spatiotemporal profile.

Since our method for spike detection pools information across channels, it is likely to obtain multiple estimates for a given spike event (i.e., s_k and s_j such that $s_k - s_j << L$). Thus, as the first step in the alignment procedure, we removed likely duplicate spike events. The magnitude of each waveform was computed as $a_k = c^{-1} \sum_{i=1}^{c} \left\| \mathbf{w}_{i,k} \right\|_2 / \sigma_i$. We then computed the difference between consecutive pairs of spikes: $\Delta_k = s_{k+1} - s_k$. For any $\Delta_k < L$, we removed s_k from S if $a_k < a_{k+1}$; otherwise, we removed s_{k+1}. This process repeated until all $\Delta_k \geq L$. We biased the removal of spike events corresponding to smaller magnitude waveforms for two reasons. For one, the smaller waveform could have been due to the spike being temporally farther from the waveform's absolute maximum, such that the extracted waveform contained more of the background EMG noise on each channel. For two, larger amplitude waveforms tend to correspond to higher threshold (and therefore less frequently observed) MUs[65,138]. (Although, we should note that this relationship is not absolute as the distance of muscle fibers from the EMG electrode also affects spike amplitude[131].)

Following the removal of duplicate spikes, remaining spike events were shifted to maximally align all W_k to each other. For two waveforms, this is a straightforward process: simply shift one of their spike indices by the lag that maximizes their cross-correlation function. For more than two waveforms, however, the process becomes dicier. A naive approach would be to randomly sample a pair of waveforms from the population and align them to one another, then iterate until convergence. Yet random sampling will inevitably select a pair of waveforms that are minimally correlated, such that aligning to maximize their cross correlation could effectively align to noise signatures in their traces. This approach could (with high likelihood) yield poor alignment, incur large shifts that prevent the algorithm from converging, and require recomputing the cross correlation for the same pairs of waveforms, which would be computationally expensive. Instead, we sought a process that was deterministic, only required computing the correlation between any pair of waveforms once, and was guaranteed to converge after a small number of iterations.

To develop some intuition for our approach, consider a case in which we have multiple noisy observations of waveforms originating from MU1 and MU2. It would not much matter for the purposes of clustering exactly how the waveforms from MU1 are aligned relative to those from MU2, so long as all waveforms from MU1 are well aligned to one another (and same for MU2). We can also assume that a waveform from M1 will look more similar to another from MU1 than to a waveform from MU2. Thus, we could envision shifting spikes in a sequence of steps. First, we shift spikes to align each waveform with its most similar match. In subsequent steps, we require that aligned waveforms are shifted in lockstep. Thus, if we repeat this process, prioritizing alignment of waveforms to their most similar match (it bringing additional waveforms along with it), we can safely align all waveforms to one another while also ensuring that waveforms belonging to a particular MU are well aligned with themselves.

To make the above concrete, let $\mathbf{W} = \{W_1, W_2, \ldots, W_N\}$. For each $W_k, W_j \in \mathbf{W}$, we computed their cross-correlation function,

$$R_{k,j}(\tau) = \sum_{i=1}^{c} (\mathbf{w}_{i,k} * \mathbf{w}_{i,j})(\tau), \quad \tau \in [-L, L]. \tag{2.9}$$

For computational efficiency, **eq.** (2.9) was evaluated for $j > k$ (because symmetry) and by pre-transforming each $\mathbf{w}_{i,k}$ into the Fourier domain so that the cross-correlation could be computed via multiplication. Our goal was to identify a series of time shifts for each spike:

$$\mathbf{\Delta}_k = \begin{bmatrix} \Delta_{k,1} & \Delta_{k,2} & \cdots & \Delta_{k,m} \end{bmatrix} \tag{2.10}$$

such that

$$s_k \mapsto s_k + \sum_{m'=1}^{m} \Delta_{k,m'} \tag{2.11}$$

maximizes the correlation between all pairs of waveforms.

The core of our alignment algorithm ("shift seek") is displayed in Listing **2.1**. In the first block, the variables `pkXC` and `optLag` are matrices with elements

$$\texttt{pkXC(k, j)} = \max_{\tau} \left| R_{k,j}(\tau) \right|, \quad \texttt{optTau(k, j)} = \operatorname{argmax}_{\tau} \left| R_{k,j}(\tau) \right|.$$

In the second block, a shift sequence is constructed for each s_k by first mapping W_k to its "nearest neighbor" (most similar other waveform). In subsequent steps, nearest neighbors are identified across groups; all waveforms within a pair of groups are mapped according to one pair of nearest neighbors across the groups. In other words, once a waveform has been mapped with another, their spikes are shifted together. This process repeats until all waveforms have been grouped together. Since this process relies on halving the number of independently "shiftable" waveforms at each iteration, the process terminates after no more than $\log_2(N)$ iterations. Note that spikes are not actually shifted during this process; each row in `shiftSeq` simply provides a sequence by which to shift each spike such that its corresponding waveform will be maximally correlated with all others.

A sequence of time lags (**eq.** (2.10)) was constructed by replacing each element in `shiftSeq` (defining a shift $k \mapsto j$) with the corresponding lag from `optLag`. Each spike was then aligned by summing over time lags, as **eq.** (2.11).

Since computing **eq.** (2.9) becomes computationally expensive for large N, we processed waveforms in batches. To increase the likelihood of similar waveforms being included in the same batch, we re-sorted the waveform arrays by a weighted sum of their norms: $\tilde{a}_k = c^{-1} \sum_{i=1}^{c} (i \cdot (1000/c)) \cdot \left\| \mathbf{w}_{i,k} \right\|_2 / \sigma_i$. Waveforms were then processed, in order, in batches no larger than 100 (performance was not noticeably different for larger batches, but runtime does increase with the square of the batch size). After the waveforms in each batch were aligned, a new batch was created using the mean waveform template from the first 100 batches and alignment repeated. This process contin-

```
%% Compute waveform cross-correlation
pkXC, optTau = wavexcorr(waveforms); % symmetric and anti-symmetric
N = size(pkXC,1);
pkXC(1:(1+N):N^2) = -Inf; % remove diagonal elements from shift candidates

%% Construct shift sequence
shiftSeq = (1:N)';
while length(unique(shiftSeq(:,end))) > 1
    shiftFrom = unique(shiftSeq(:,end));
    shiftTo = shiftFrom;
    [~,bestMatch] = max(pkXC(shiftTo,shiftTo),[],2);
    for ii = 1:length(shiftTo)
        shiftTo(ii) = shiftTo(bestMatch(ii));
    end
    [~,mapIdx] = ismember(shiftSeq(:,end),shiftFrom);
    shiftSeq = [shiftSeq, shiftTo(mapIdx)];
end
```

Listing 2.1: Shift seek alignment algorithm (MATLAB)

ued until all waveforms were aligned.

Clustering To cluster the aligned waveforms, we adapted the density-based algorithm (ISO-SPLIT) used by the MountainSort package[119,139]. ISO-SPLIT leverages the observation that clusters in an N-dimensional space can be identified by projecting the data onto a one-dimensional (1D) axis and testing whether the resulting distribution is unimodal. One challenge of the approach requires knowing how to find the optimal projection axis. To tackle this problem, the authors adopt a bottom-up, or agglomerative approach: the data are over clustered (using k-means with an excessively large k), and the projection axis is chosen as that which connects the centroids of a pair of clusters. At each iteration, a pair of clusters are either merged (if the distribution of 1D projections is unimodal) or split (if the distribution is not unimodal). This process repeats until the cluster assignments stabilize. Some drawbacks of this approach are that it requires a sufficiently large initial choice of k, so that all true clusters are separated; there is no way to know how large k needs to be for a given data set; and since clusters can be split or merged at each iteration, there is also no way of knowing how long it will take for the algorithm to converge.

Instead, we adopted a top-down approach, which we describe generally first, then specifically in

the next paragraph. Independent components of the data matrix were identified using Hyvärinen's stabilized fixed-point algorithm with a tanh contrast function[140]. Each independent component was used as a projection vector for the ISO-SPLIT algorithm[139]. ISO-SPLIT uses Hartigan's Dip Test as a statistical test of unimodality[141]. For any projection vectors for which the null hypothesis (that the distribution of 1D projections is unimodal) was rejected with $p < 0.001$, we kept the vector that produced the largest separation and split the labels along the optimal split point (where the spacings between the ordered 1D projections is largest). If the null hypothesis was not rejected, then the cluster labels were preserved. This process repeated until all clusters were deemed unimodal. We refer to this algorithm as DCAPP (divisive cluster analysis via projection pursuit).

Waveforms were clustered in two steps. In the first step, we initialized all waveforms with the same label. Labels were then split using single-channel data. For each channel, we constructed $Y_i = \begin{bmatrix} \mathbf{w}_{i,1}^\top & \mathbf{w}_{i,2}^\top & \cdots & \mathbf{w}_{i,N}^\top \end{bmatrix}^\top \in \mathbb{R}^{N \times L}$. Y_i was projected onto its first two principal components (PCs) and each row augmented with $\|\mathbf{w}_{i,k}\|_2$. This transformed $Y_i \mapsto Z_i \in \mathbb{R}^{N \times 3}$. Label assignments were updated using DCAPP to cluster Z_i. This process was repeated for each channel, carrying over the updated labels to each channel. Thus, each channel afforded the opportunity to further split the existing clusters. In the second step, we used the multi-channel waveforms to provide one final opportunity to split clusters. We constructed $Y \in \mathbb{R}^{N \times (cL)}$, where each row contained each waveform's single-channel traces stacked end to end. Y was projected onto its top three PCs, and the resulting set of features and existing label assignments passed to DCAPP for further refinement. By design, this approach over splits the waveforms. However, in contrast with ISO-SPLIT, our approach does not require pre-specifying how many clusters to split. Furthermore, in this stage, labels were only split, but never merged, which reduces runtime. We resolved spuriously split clusters in the next stage of the pipeline.

Merging We used relatively short waveform durations for alignment and clustering ($L = 3$ ms). This overly tight window was chosen to reduce the influence of noise and only focus on the most

salient waveform features in the preceding stages. After the clustering stage, we re-extracted waveforms using a longer window (L = 10 ms), which served as the final template duration.

Using the updated waveforms, we searched for spuriously split clusters to merge. We computed the similarity (maximum absolute cross-correlation) between the waveform templates (mean over observations) for each pair of clusters. Working from most to least similar pairs of templates, if their similarity exceeded a threshold (0.7), then the waveforms from both clusters were pooled, projected into their first three PCs and the resulting set of features passed to DCAPP. If the feature distribution was determined to be unimodal, then the cluster labels were merged and the waveform template similarities updated. Otherwise, the cluster labels were preserved. This process repeated until all pairs of waveform templates failed to exceed the similarity threshold or were identified as multi-modal by DCAPP. Thus, in this step, clusters were either merged or no action was taken.

Triaging By this stage, we will have obtained a set of waveform clusters, each putatively belonging to one MU. It is likely, however, that each cluster contains outliers – noisy observations of the stereotyped MU action potential waveform, either due to improper cluster assignments or superpositions with waveforms belonging to different MUs. We therefore triaged outliers to refine our estimate of each waveform template. Any cluster containing fewer than 5 observations was considered noise and discarded. For each remaining cluster, we projected its waveforms onto their first three PCs and binned the features in a 3D grid using 100 equally spaced bins per dimension. Bin counts were smoothed with a 3D Gaussian blur (`imgaussfilt3`, MATLAB) to estimate the density of each observation in the feature space. We then removed the bottom quintile of observations based on their estimated density.

Noise covariance estimation In subsequent stages, we deconvolved EMG signals with our waveform templates. Deconvolution algorithms often require an estimate of the spatiotemporal noise

covariance. We estimated the noise covariance as a block Toeplitz matrix[142]:

$$\Lambda_0 = \begin{bmatrix} \mathrm{E}\left[\boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top\right] & \mathrm{E}\left[\boldsymbol{\eta}_1\boldsymbol{\eta}_2^\top\right] & \cdots & \mathrm{E}\left[\boldsymbol{\eta}_1\boldsymbol{\eta}_c^\top\right] \\ \mathrm{E}\left[\boldsymbol{\eta}_2\boldsymbol{\eta}_1^\top\right] & \mathrm{E}\left[\boldsymbol{\eta}_2\boldsymbol{\eta}_2^\top\right] & \cdots & \mathrm{E}\left[\boldsymbol{\eta}_2\boldsymbol{\eta}_c^\top\right] \\ & & \vdots & \\ \mathrm{E}\left[\boldsymbol{\eta}_c\boldsymbol{\eta}_1^\top\right] & \mathrm{E}\left[\boldsymbol{\eta}_c\boldsymbol{\eta}_2^\top\right] & \cdots & \mathrm{E}\left[\boldsymbol{\eta}_c\boldsymbol{\eta}_c^\top\right] \end{bmatrix} \tag{2.12}$$

where $\mathrm{E}\left[\boldsymbol{\eta}_i\boldsymbol{\eta}_j^\top\right]$ is the auto-/cross-correlation matrix (a Toeplitz matrix) for noise segments from channels i and j.

The $\boldsymbol{\eta}_i$ were extracted from the lowest energy segments in the EMG signals. To identify these segments, we split the EMG signals into bins whose edges were determined by a 500 ms buffer around each initially detected spike. Within each bin, we computed the average (over samples and channels) of the signal energy on each channel (normalized by the noise standard deviation, σ_i). The multi-channel signals were then re-ordered based on their mean energy. We set $\boldsymbol{\eta}_i$ as the first 1 s segment of the re-ordered signals from the i^{th} channel.

Auto-curation We automatically curated our set of templates by running the deconvolution algorithm on short segments of data and removing MUs that exhibited irregular spiking statistics, caused the residual energy in the reconstructed segments to increase, or were simply unused. We implemented deconvolution using Bayes optimal template matching (BOTM) with subtractive interference cancellation (SIC) to resolve instances of overlapping waveforms[118].

In each curation round, 30 random 10-s long segments of data (30 "trials") were deconvolved. One MU was discarded if it met any of following criteria (ordered by priority): (1) its median (across trials) interspike interval coefficient of variation exceeded 1, (2) the ratio of its residual energy to the energy of the raw EMG signal exceeded 1.1 on any channel for 10% of trials, or (3) no spikes were detected on any trial. If multiple MUs met one of the thresholds, then the worst

offender was removed. This process repeated until no MUs met any of the removal thresholds.

Manual curation Following automatic curation, templates were inspected by hand using a custom MATLAB GUI to split, merge, or remove clusters as needed.

Deconvolution The full set of EMG signals were deconvolved with the remaining waveform templates using BOTM and SIC to obtain the final spike times for each MU.

2.6.2 Optimal Motor Unit Recruitment

Introduction

What determines the manner in which MUs are recruited and coordinated to produce an output? In certain circumstances, Henneman's size principle reflects the computationally optimal recruitment strategy. Neural control signals are corrupted by 'signal-dependent noise' (SDN) whose standard deviation increases linearly with the mean of the signal[143]. The accumulation of SDN over the course of a movement incurs undesirable variability in the final position. Voluntary forces generated via isometric muscle contractions are characterized by SDN, which is not observed during neuromuscular stimulation, suggesting that SDN arises from volitional MU recruitment, rather than peripheral sources[109]. This likely relates to large MUs requiring bigger input currents than smaller MUs to reach their critical firing (recruiment) threshold[45,64,144]. As a computational proof of principle, models of isometric force production in which MU recruitment follows the size principle (small MUs recruited before larger MUs) produce less noise than models in which the recruitment order is reversed or randomized[109]. The size principle can therefore be understood as the minimum-variance recruitment strategy[110].

Though static force production is most optimally achieved via the size principle, it is unclear whether it remains optimal when forces need to change quickly. MUs are morphologically and physiologically heterogeneous, varying in both size and speed. MU size conventionally refers to

its innervation ratio (the number of muscle fibers innervated by the α-motoneuron)[46], whereas speed relates to the contractile properties of its muscle fibers[104]. The size of a MU determines its maximal force capacity[47] and its speed determines the activation and relaxation dynamics of its muscle fibers[145]. Fiber activation (and, particularly deactivation) dynamics negligibly impact muscle performance during slow tasks, but limit muscle power in rapid tasks[146], most substantially in muscles composed primarily of slowly contracting muscle fibers[147]. These findings suggest that the contraction dynamics of a muscle should match the frequency of the movement that it propels and some evidence indicates that coordination across muscles is influenced by movement speed[148,149]. It has further been suggested that this principle ought to apply at the level of MU recruitment[87,89], and our empirical recordings demonstrate that flexible recruitment does occur during dynamic force production. Yet this hypothesis has not been explored computationally. More generally, it remains unclear how MU size and speed properties might interact such that the size principle is the optimal MU recruitment strategy in some circumstances, but not in others. Here, we use an model of muscle force generated by an idealized motor pool, including heterogeneity in MU size and speed, to derive optimal recruitment strategies for a broad range of muscle force profiles.

Muscle model

An action potential emitted by an α-motoneuron causes all of its innervated fibers to concurrently contract, causing a brief rise and fall in muscle tension. This muscle-tension response to a motoneuronal impulse is called the MU's twitch response, which can be modeled as[49]

$$h_i(t) = \frac{P_i}{T_i} t e^{1-t/T_i}, \quad i = 1, 2, \ldots, N \tag{2.13}$$

where P_i and T_i are the peak tension (maximum force generated by the twitch) and contraction time (latency from impulse to peak tension) and i indexes over a motor pool of N MUs. We expressed

the force generated by repeated discharges of the i^{th} MU as[150]

$$f_i(t) = (h_i * s_i)(t) = \sum_{j=1}^{k} h_i(t - t_j) \tag{2.14}$$

where $s_i(t)$ indicates the motoneural response function, which is one at times t_j and zero otherwise. Over multiple trials of the same behavior, the average force is given by

$$\bar{f}_i(t) = (h_i * r_i)(t) = \int_{t_i}^{t_f} \mathrm{d}\tau \; h_i(t - \tau) r_i(\tau) \tag{2.15}$$

where $r_i(t)$ is the trial-averaged firing rate and t_i and t_f indicate the beginning and end of the trial. We then take the trial-averaged force produced by the motor neuron pool (MNP) as

$$f_{\text{MNP}}(t) = \sum_{i=1}^{N} \bar{f}_i(t). \tag{2.16}$$

Cost functional

We would like to determine the set of MU firing rates $R = \{r_1(t), r_2(t), \ldots, r_N(t)\}$ that optimally produces a particular force profile, $\hat{f}(t)$. We define the optimal R as that which minimizes the mean-squared error between the motor pool and desired forces,

$$J(R) = \mathrm{E}\left[\left(f_{\text{MNP}}(t) - \hat{f}(t)\right)^2\right] \tag{2.17}$$

which can be decomposed as[151]

$$\begin{aligned} &= \left(\mathrm{E}\left[f_{\text{MNP}}(t)\right] - \hat{f}(t)\right)^2 + \mathrm{E}\left[\left(f_{\text{MNP}}(t) - \mathrm{E}\left[f_{\text{MNP}}(t)\right]\right)^2\right] \\ &= \mathrm{Bias}^2\left[f_{\text{MNP}}(t)\right] + \mathrm{Var}\left[f_{\text{MNP}}(t)\right]. \end{aligned} \tag{2.18}$$

Eq. (2.18) captures the fundamental trade-off in force accuracy and precision that principally determines the qualitative form of the solutions. We also imposed additional regularization constraints. First, we enforced non-negativity in the firing rates. Second, we added to the cost

$$\int_{t_i}^{t_f} \mathrm{d}t \; \lambda_m \sum_{i=1}^{N} P_i r_i^2(t) + \lambda_d \sum_{i=1}^{N} \left(\frac{\mathrm{d}r_i(t)}{\mathrm{d}t} \right)^2 \tag{2.19}$$

where $P_i = h_i(T_i)$. The first term discourages an over reliance on any particular unit by penalizing muscle contractions via the weighting P_i and the second term imposes smoothness in the firing rates; λ_m and λ_d are free parameters.

To combine **eqs.** (2.18) and (2.19), we assume that the $s_i(t)$ arise from independent inhomogeneous Poisson processes. Then, $\mathrm{Var}\left[f_{\mathrm{MNP}}(t)\right] = \sum_{i=1}^{N} (h_i^2 * r_i)(t)$ and we have

$$J(R) = \int_{t_i}^{t_f} \mathrm{d}t \; \left(f_{\mathrm{MNP}}(t) - \hat{f}(t) \right)^2 + \sum_{i=1}^{N} (h_i^2 * r_i)(t) + \lambda_m \sum_{i=1}^{N} P_i r_i^2(t) + \lambda_d \sum_{i=1}^{N} \left(\frac{\mathrm{d}r_i(t)}{\mathrm{d}t} \right)^2, \tag{2.20}$$

$$r_i(t) \geq 0 \; \forall \; t \in [t_i, t_f].$$

The non-negativity constraint in **eq.** (2.20) makes obtaining an analytic solution intractable. However, we can readily obtain numerical solutions by re-expressing the cost in a more convenient form. First, we fully expand **eq.** (2.20) and drop any terms not depending on R. This gives

$$J(R) = \int_{t_i}^{t_f} \mathrm{d}t \; f_{\mathrm{MNP}}^2(t) - 2\hat{f}(t) f_{\mathrm{MNP}}(t) + \sum_{i=1}^{N} (h_i^2 * r_i)(t) + \lambda_m \sum_{i=1}^{N} P_i r_i^2(t) + \lambda_d \sum_{i=1}^{N} \left(\frac{\mathrm{d}r_i(t)}{\mathrm{d}t} \right)^2. \tag{2.21}$$

Next, we re-write **eq.** (2.21) in matrix form. For this, note that the convolution can be expressed in

terms of matrix multiplication:

$$(h_i * r_i)(t) = \int_{t_i}^{t_f} \mathrm{d}\tau \; h_i(t-\tau) r_i(\tau) = \begin{bmatrix} h_i(0) & 0 & \cdots & 0 \\ h_i(\Delta t) & h_i(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_i(t_f) & h_i(t_f - \Delta t) & \cdots & h_i(0) \end{bmatrix} \begin{bmatrix} r_i(t_i) \\ r_i(t_i + \Delta t) \\ \vdots \\ r_i(t_f) \end{bmatrix} = H_i \mathbf{r}_i$$

where $H_i \in \mathbb{R}^{T \times T}$ is a Toeplitz matrix and $T = t_f - t_i$. Then

$$f_{\text{MNP}}(t) = \sum_{i=1}^{N} (h_i * r_i)(t) = \begin{bmatrix} H_1 & H_2 & \cdots & H_N \end{bmatrix} \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_N \end{bmatrix} = \widetilde{H} R \tag{2.22}$$

which provides

$$\hat{f}(t) f_{\text{MNP}}(t) = \hat{\mathbf{f}}^\top \widetilde{H} R \tag{2.23}$$

where

$$\hat{\mathbf{f}} = \left[\hat{f}(t_i),\, \hat{f}(t_i + \Delta t),\, \ldots,\, \hat{f}(t_f) \right]^\top.$$

The derivative can also be implemented via matrix multiplication:

$$\begin{aligned} \frac{\mathrm{d}r_i(t)}{\mathrm{d}t} &= \left[r_i(t_i + \Delta t) - r_i(t_i),\, r_i(t_i + 2\Delta t) - r_i(t_i + \Delta t),\, \ldots,\, r_i(t_f) - r_i(t_f - \Delta t),\, 0 \right]^\top \\ &= \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \begin{bmatrix} r_i(t_i) \\ r_i(t_i + \Delta t) \\ \vdots \\ r_i(t_f) \end{bmatrix} = D \mathbf{r}_i \end{aligned} \tag{2.24}$$

where the final row of zeros in D ensures that $D\mathbf{r}_i$ has dimensions $T \times 1$. To apply **eq.** (2.24) for

all N firing rate functions, we let

$$\widetilde{D} = \mathbb{1}_N \otimes (\lambda_d D) \tag{2.25}$$

where $\mathbb{1}_N$ is the $N \times N$ identity matrix and $\otimes$ denotes the Kronecker product. Similarly, let $A \in \mathbb{R}^{N \times N}$ denote a matrix with entries $a_{ij} = \delta_{ij} P_i$, where δ_{ij} is the Kronecker delta. Then let

$$\widetilde{A} = (\lambda_m A) \otimes \mathbb{1}_T. \tag{2.26}$$

Using **eqs.** (2.22), (2.23), (2.25) and (2.26), **eq.** (2.21) can be rewritten as

$$\begin{aligned} J(R) &= R^\top \widetilde{H}^\top \widetilde{H} R - 2\hat{\mathbf{f}}^\top \widetilde{H} R + (\widetilde{H} \odot \widetilde{H}) R + R^\top \widetilde{A} R + R^\top \widetilde{D}^\top \widetilde{D} R \\ &= R^\top \left(\widetilde{H}^\top \widetilde{H} + \widetilde{A} + \widetilde{D}^\top \widetilde{D} \right) R + \left((\widetilde{H} \odot \widetilde{H}) - 2\hat{\mathbf{f}}^\top \widetilde{H} \right) R, \quad R \geq 0. \end{aligned} \tag{2.27}$$

where $\odot$ is the Hadamard product.

Eq. (2.27) has the exact form of a quadratic programming problem, for which the goal is to solve

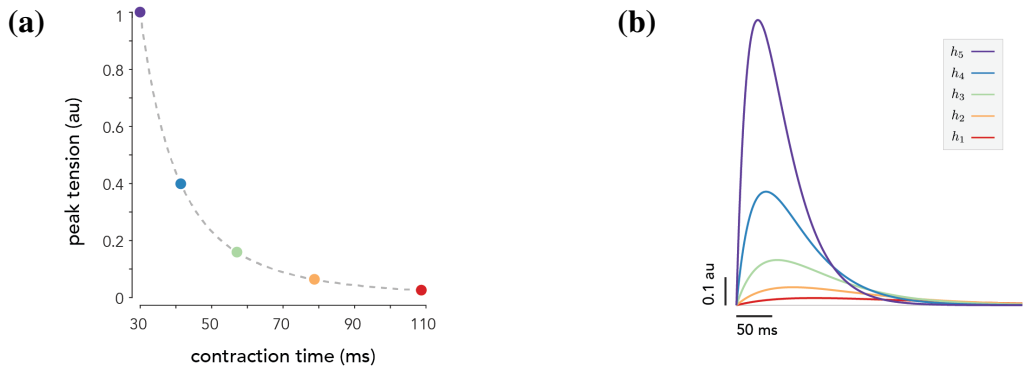
$$\min_{\mathbf{x}} \ \mathbf{x}^\top Q \mathbf{x} + \ell^\top \mathbf{x}$$

with or without bounds and constraints on $\mathbf{x}$. Quadratic programming is a standard class of optimization problems, for which various solution methods have been developed. Thus, we can readily obtain numerical solutions to **eq.** (2.20) using the form given in **eq.** (2.27).

Results

We considered a motor neuron pool (MNP) containing five MUs. Their twitch responses were simulated using **eq.** (2.13) with

$$P_i = e^{i * \ln(PR)/N}, \quad T_i = T_L \left(\frac{1}{P_i} \right)^{\ln(TR)/\ln(PR)} \tag{2.28}$$

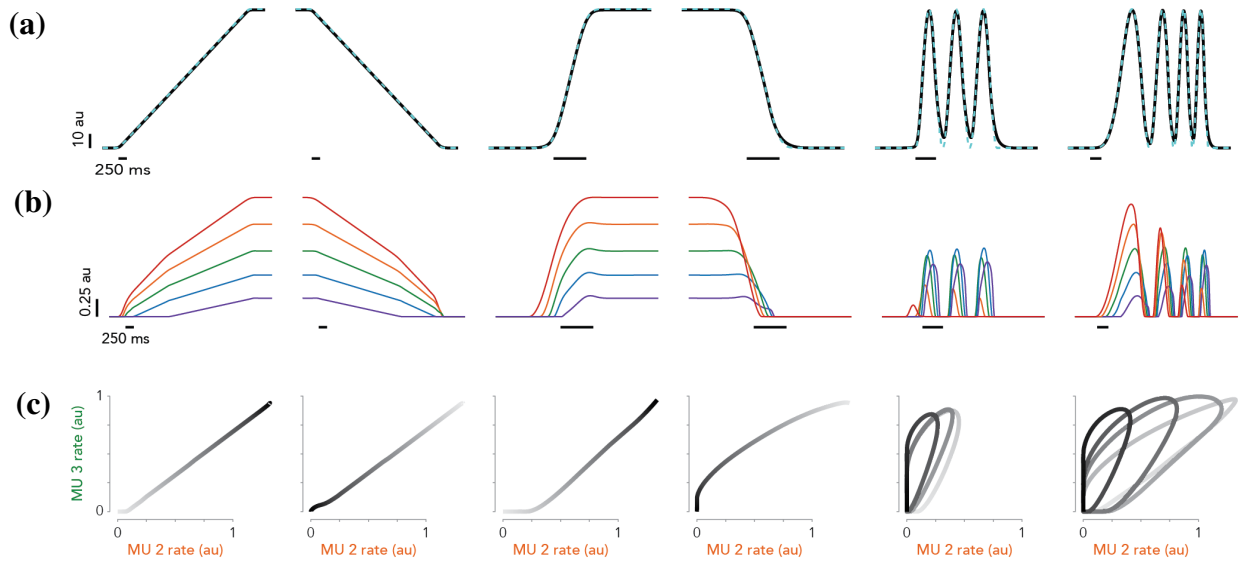


Supplementary Figure S2.11: Twitch responses. (**a**) Twitch parameters for each simulated MU. **(b)** Simulated twitch responses (h_i). Traces are colored to match the set of parameters in ***a*** used to generate them.

as described by Fuglevand *et al.*[49], where T_L denotes the maximum contraction time, PR denotes the range of peak tensions, and TR denotes the range of contraction times. We set $T_L = 150$, $PR = 100$, and $TR = 5$, in accordance with empirically measured twitch responses[51]. The twitch responses and their parameters are shown in **fig. S2.11**.

We numerically derived the optimal firing rates that minimized the mean-squared error between the MNP force and a particular force profile. We considered several force profiles that were employed during experiments: slow (4 s) and fast (250 ms) increasing and decreasing ramps, a sinusoid with constant frequency (3 Hz) and another with linearly increasing frequency (0-3 Hz). Solutions were obtained using `quadprog` (MATLAB) to solve **eq.** (2.27). The target and model-generated force profiles are shown in **fig. S2.12a**, indicating that the model accurately matched the target force over the range of profiles considered.

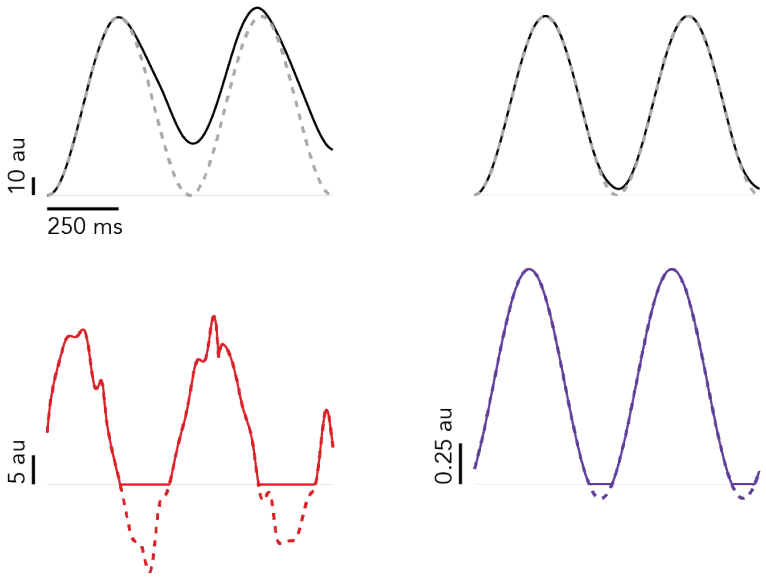
The predicted firing rates are shown in **fig. S2.12b**. For slow (4 s ramp) force profiles, the model predicted that MUs should be recruited in accordance with the size principle. Namely, the MU with the smallest/slowest twitch response (**fig. S2.12b**, *red traces*) was recruited first and all others were successively recruited in order by the amplitude of their twitch responses, ending with the largest MU (*violet traces*). The model also recruited MUs in accordance with the size principle



Supplementary Figure S2.12: Optimization predictions. (**a**) Target forces ($\hat{f}(t)$; *cyan traces*) and simulated motor pool force ($f_{\text{MNP}}(t)$; *black traces*). (**b**) Predicted firing rates for each unit (solutions to **eq.** (2.20)). Trace colors match that of the MU twitch responses shown in **fig. S2.11b** (e.g., red corresponds to the MU with the smallest/slowest twitch response). (**c**) Predicted MU 3 rate plotted against MU 2. Traces are shaded from light to dark to indicate the progression of time.

for a fast (250 ms) increasing ramp force, but leveraged more flexible strategies for other target forces. For the fast decreasing ramp force (**fig. S2.12b**, *fourth column from the left*), the smallest MU was de-recruited first and the largest MU de-recruited last, immediately prior to force offset. This contrasts with the classical "first-in-last-out" recruitment ordering employed by the model for the slow ramp forces as well as empirically observed MU activity patterns during steady force production[7]. Generating rapidly oscillating forces also involved flexible MU recruitment. During the 3 Hz sinusoidal force (**fig. S2.12b**, *fifth column*), the smallest MU (*red*) was essentially not used at all, while the largest MU (*violet*) was roughly three times more active than predicted during a slowly increasing ramp force. Gradual shifts in recruitment were predicted as force frequency linearly increased (**fig. S2.12b**, *right-most column*), with the smallest MUs becoming less active as the largest MUs became more active.

We used the state space view to visualize pairwise MU activity patterns. **Fig. S2.12c** shows the predicted firing rate of MU 3 versus MU 2 for each force profile (corresponding to the *green* and



Supplementary Figure S2.13: Recruiting fast MUs increases force accuracy. The firing rate for one MU, $r_i(t)$, was obtained by deconvolving $(h_i * r_i)(t) = \hat{f}(t)$, where $\hat{f}(t)$ was a 2 Hz sinusoid (*dashed traces, top row*). The directly obtained firing rates are the *dashed traces, bottom row*. *Left* corresponds to MU1 (slowest/smallest) and *right* corresponds to MU5 (fastest/largest). To correct for the moments when firing rates are negative, they can simply be rectified (*solid traces, bottom row*). Convolving the rectified firing rates with each MU's twitch response generates new forces (*solid traces, top row*). The new force is more accurate for MU5 (faster) than MU1 (slower).

orange traces in **fig. S2.12b**, respectively). As described in [**main text**], their activity trajectory traversed a 1-dimensional monotonic manifold, as expected, for steady forces. Modest departures from that manifold were observed during rapid force offset, corresponding to the reversal in the classical de-recruitment order, with more dramatic departures observed for sinusoidal force profiles.

Intuitively, the predicted changes in recruitment strategy during rapid force offset arises from the need to recruit MUs whose activation-deactivation dynamics (speed of twitch force rise and fall) match the dynamics of the generated force[152]. To demonstrate why this affects the model predictions, we considered generating a high-frequency (2 Hz) sinusoidal force with one of two MUs: the slowest (MU1) or fastest (MU5) in our model. That is, we sought to obtain $r_i(t)$ such that $(h_i * r_i)(t) = \hat{f}(t)$. In an ideal world, $r_i(t)$ can be straightforwardly obtained by deconvolving the target force with the twitch response. These solutions are shown in **fig. S2.13** (*bottom, dashed*

traces); convolving each solution with its corresponding twitch response reproduced the intended sinusoid (*top, dashed traces*), as expected. However, these optimal firing rates involved moments when the functions became negative to compensate for the twitch deactivation dynamics. This phenomenon occurred for both MUs, but was largest for MU1 (**fig. S2.13**, *left*), which had a slower twitch response. Since firing rates can not, of course, become negative, a naive solution would be to rectify each firing rate function (**fig. S2.13**, *bottom, solid traces*). When the rectified firing rates were then convolved with the twitch responses, the force generated by MU5 was largely unchanged (**fig. S2.13**, *top right, solid trace*), but deviated much more dramatically from the intended force profile for MU1 (*top left, solid trace*). Thus, generating rapid forces with fast MUs maximizes force accuracy.

Optimal MU recruitment, as predicted by our model, fundamentally involves a trade off in bias and variance (**eq.** (2.18)). When forces change slowly, the optimal strategy is to rely on small MUs, since they incur less variance due to signal-dependent noise[109]. In these situations, MU activation-dynamics negligibly impact incurred bias. On the other hand, naively relying on small MUs in the same manner when forces need to quickly terminate can incur substantial mean errors, as demonstrated in **fig. S2.13**. These errors can be avoided to a certain degree by phase advancing the activation of slow MUs, but when forces need to change rapidly enough, the only feasible strategy is to recruit and control MUs more flexibly.

Chapter 3: High-dimensional neural manifolds for complex muscle control

The dimensionality of a neural population – the 'degrees of neural freedom' – reflects the complexity of internal computations as well as the number of signals that can be 'read out' downstream. Recent studies of primary motor cortex (M1) have suggested that neural activity resides in a low-dimensional manifold that is preserved across different tasks[26,32,35,36,123]. On the other hand, we recently reported hundreds of neural modes in M1 during an isometric force-tracking task using a cross-trial reliability measure (**chapter 2**). Here we investigate the reason for this high dimensionality. We find a reorganization of neural activity patterns across force conditions, similar to the reorganization observed across preparatory and movement epochs during reaching[32]. In particular, neural activity unfolded in orthogonal subspaces during the generation of static forces and rapidly oscillating force profiles. These findings indicate that the high dimensionality observed in M1 indicates that unrelated neural activity modes are used to generate different behaviors.

3.1 Introduction

Animals are capable of flexibly generating different movements. Early efforts to elucidate the neural mechanisms of movement considered a relatively restricted range of behaviors. Evarts pioneered investigations of neural activity in the primary motor cortex (M1) of monkeys performing alternating flexion-extension movements of the wrist[8] or generating static torques against an external load[9]. During these tasks, single-cell responses vary primarily with the pattern of muscle activation and the direction or rate of change of exerted force. In similar tasks involving one-dimensional control of a manipulandum, neural activity in M1 correlates well with muscle activity[13], force or its derivative[10,11], joint position or intended movement direction[12,14]. It was suggested that the usage of simplistic, one-dimensional tasks might limit the understanding of cortical control of multiple degree-of-freedom joints in two or three dimensions[17]. Yet during ballistic reaches[15–19] or continuous arm movements[20], neural activity patterns remained largely understood as encoding kinematic or kinetic parameters.

The conceptual and mathematical models of motor cortex have predominantly shifted from a representational to a dynamical systems perspective[25], based on an understanding that the goals of motor cortex are to plan, generate and control movement[2,153,154]. Pursuant to its goals, motor cortex constitutes a dynamical system that governs the time evolution of neural activity[24,33]. Neural dynamics in motor cortex exhibit several robust features across different species and tasks. First, during arm reaches[24,26–30,34,155] and pedaling[31], the largest signals in the population response display a condition-invariant translation and/or strong rotations. Rotational structure does not trivially arise from correlations across times, neurons or conditions[156]. Second, motor cortical population activity has the geometric property of low 'trajectory tangling', meaning that a particular activity pattern never leads to multiple dissimilar states[31]. Relative to motor cortex, tangling is dramatically higher in visual cortex, somatosensory cortex, and electromyographic (EMG) signals recorded from arm muscles[31]. The disparity in tangling across areas may reflect that motor cor-

tex can be well approximated as a dynamical system whose computations principally rely on its internal recurrence, rather than external inputs[157]. Third, preparatory activity unfolds in a neural subspace orthogonal to that occupied for movement execution[26,32]. Stated more simply, neurons that co-modulate as one prepares to move are completely unrelated as movement unfolds. Despite this reorganization in the neural covariance structure across temporal epochs within a movement, it does not appear to be the case that similar reorganization occurs across movements.

How motor cortex flexibly produces different behaviors remains largely unclear. On one hand, it could rely on completely different computations to generate different movements; on the other, it could simply reuse a relatively small number of neural modes. For example, neurons whose activity tend to increase together during one behavior might decrease together for another behavior – the individual activity patterns differ across behaviors, while preserving the neural covariance structure. The available evidence favors the latter interpretation. Monkeys trained to control virtual cursor movements by modulating neural activity mapped to a 10-dimensional manifold can quickly adapt to within-manifold perturbations of the decoder (i.e., preserving the neural covariance structure), but fail to learn off-manifold perturbations[36]. And relatively few signals (10-20) are required to explain most of the variance in hundreds of motor cortical neurons during one or more behaviors[26,32,35,36]. The ability of a small number of signals to explain most of the variance in a large population of neurons appears not to be unique to the motor system, as similar findings are observed in somatosensory, olfactory, visual, prefrontal, and hippocampal systems[6,37]. In light of these observations, it has been recently proposed that motor cortex generates movement by activating a few neural modes confined to a low-dimensional manifold, thereby simplifying the control problem[6]. Though computationally appealing, the simplicity observed in the neural control of movement could instead reflect simplicity in the behavioral tasks used by experimenters, rather than a fundamental property of motor cortex[37].

Recent evidence indicates the presence of hundreds of neural modes in M1 during an isomet-

ric force-tracking task (**chapter 2**). This could have one of two interpretations: either many of the observed dimensions are incidentally evoked by the task, or the task taps into a previously unexplored regime in motor control, wherein a large number of signals are leveraged to generate fundamentally different behaviors. Here we investigate these possibilities, revealing evidence that favors the latter interpretation. We observe not only a large number of neural modes, but also that neural activity unfolds in orthogonal subspaces for the most dissimilar behaviors.

3.2 Results

3.2.1 Task and behavior

A rhesus macaque generated isometric forces to modulate the vertical position of a cursor and intercept a scrolling dot path (**fig. 3.1a**). Each behavioral condition corresponded to one of twelve instructed force profiles: three static forces, over a range of amplitudes; four linearly ramping forces, increasing or decreasing, quickly (250 ms) or slowly (4 s); four sinusoidal forces with constant frequency (0.25, 1, 2, and 3 Hz); and one sinusoidal force whose frequency linearly increased from 0 to 3 Hz (a "chirp" signal).

We recorded from 1257 neurons in M1. Most neurons (881) were recorded using 128-channel Neuropixels probes (**fig. 3.1b**) and the remainder with 32-channel linear probes. The responses of multiple MUs were isolated from multi-channel EMG signals recorded from the deltoid and

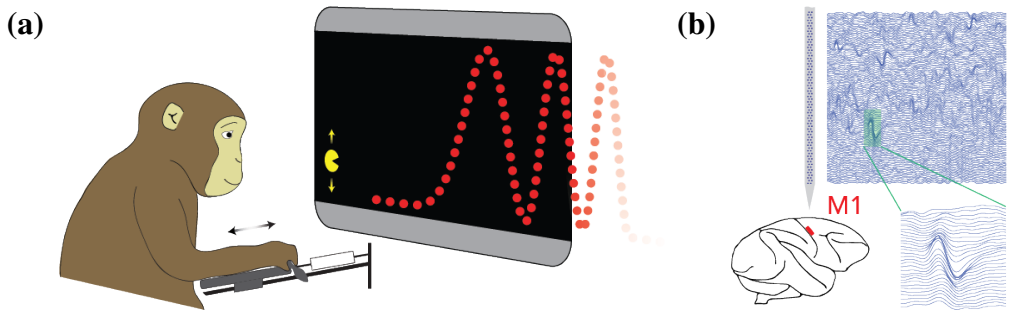


Figure 3.1: Task and neural recordings. (**a**) A monkey modulated the force generated against a load cell to control the vertical position of Pac-Man and intercept a scrolling dot path. (**b**) Neuropixels probes were used to record neural activity in M1.

triceps, as described previously (**chapter 2**). In three sessions, Neuropixels recordings were conducted simultaneously with EMG recordings from the lateral head of the triceps. In a separate session, we recorded EMG signals from eight different muscles of the upper arm (not concurrently with neural recordings).

EMG signals generally resembled the force profiles, albeit with some differences across muscles and conditions (**fig. S3.9**). The lateral triceps correlated most strongly with force across all conditions. All muscles were maximally activated during sinusoidal conditions, yet some were preferentially recruited for moderate (1 Hz) or high (3 Hz) frequency forces. Similar differences were observed within a condition. For example, the deltoids were primarily activated during the first and last phase of the chirp force, whereas the triceps were most active during the final phases.

3.2.2 Single-neuron response features

The response features of some neurons were consistent with those observed during simple and complex tasks. The firing rate of some neurons resembled force (**fig. 3.2a**) or the rate of change of force, as observed during flexion-extension tasks[8,9]. For example, unit 1183 tracked force amplitude across all conditions (**fig. 3.2b**); unit 341 varied inversely, but reliably with force amplitude (**fig. 3.2c**); and unit 862 was minimally modulated during steady force conditions, but discharged transiently when force changed quickly (**fig. 3.2d**). Many neurons also displayed consistent, multiphasic responses, as observed during cycling[31]. For example, units 1183 and 862 displayed double-peaked responses during each period of the 2 Hz sinusoidal condition (**figs. 3.2b** and **3.2d**). Consistent response features could also be observed across conditions. For example, the response of unit 443 during the first period of the chirp condition closely resembled its response during the first period of the 1 Hz sinusoid, whereas its response during the final (higher frequency) periods of the chirp condition better matched the 3 Hz sinusoid (**fig. 3.2e**).

Most neurons, however, did not bear a discernibly direct relationship with the behavior, particu-

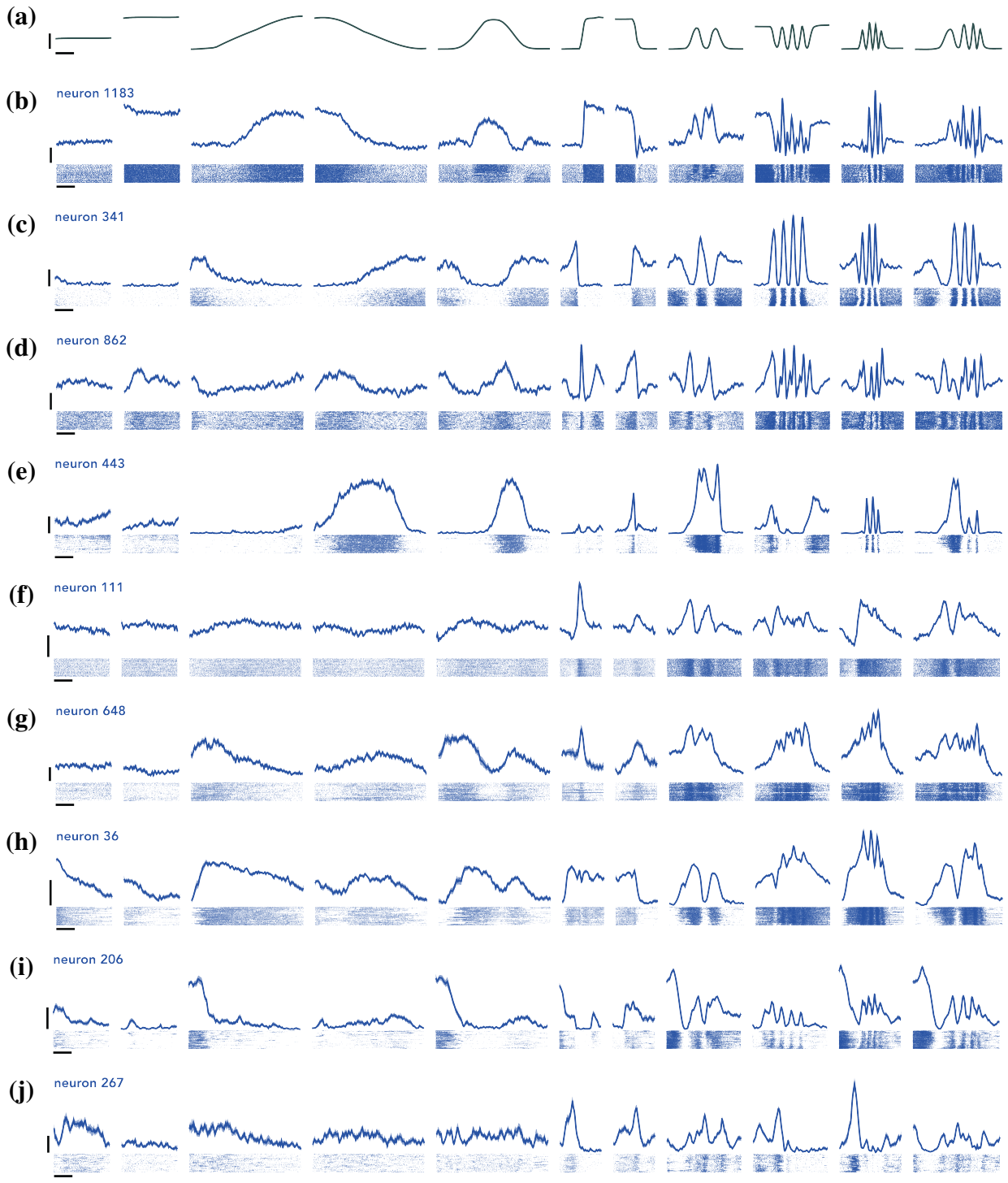


Figure 3.2: Example primary motor cortex (M1) neuron responses. (**a**) Trial-averaged forces for 11 of 12 conditions (intermediate static force condition is omitted for space). Vertical scale bar indicate 8 N. Horizontal scale bar indicates 1 s. (**b**-**j**) Trial-averaged firing rates with standard error (*top*) and single-trial spike rasters (*botom*). Vertical scale bars indicate 20 spikes/s. Horizontal scale bars indicate 1 s. (reproduced from **fig. S2.10**)

larly when comparing across multiple conditions. For example, units 443, 111, and 648 discharged transiently during a rapid ramp force (**figs. 3.2e** to **3.2g**), which might suggest that they encoded the rate of change of force. However, relative to its response during a fast decreasing ramp, unit 443 was more active during a slow decreasing ramp, not at all active when force oscillated at 2 Hz, but comparably active when force oscillated at 3 Hz (**fig. 3.2e**). Similarly, whereas units 111 and 648 both discharged transiently during a fast increasing ramp, their responses predominantly linearly decreased or increased when force oscillated at 3 Hz (**figs. 3.2f** and **3.2g**). As additional examples, the firing rate of unit 36 resembled force amplitude during low frequency (1 Hz) oscillations, but varied linearly over a large range ($\sim$ 40 spikes/s) when force was maintained statically at a low amplitude (**fig. 3.2h**); and units 206 and 267 exhibited putative preparatory activity in certain conditions, but not others (**figs. 3.2i** and **3.2j**).

3.2.3 Neural correlates with behavior

Previous efforts to relate motor cortical activity to the motor output used extrinsic movement parameters, such as force[8–11,13] or limb kinematics[15–20]. This representational view of motor cortex has largely fallen out of favor[25], but since α-motoneurons constitute the final neural layer in the motor pathway, relating M1 activity with that of motoneurons could provide a more direct means of determining whether motor cortex merely encodes motor output. To investigate and quantify the relationship between motor cortex and behavior, we fit single-neuron responses to force or motor unit activities. The firing rate of each neuron was fit to an encoder model of the form[22] $r_i(t-\tau) \sim \boldsymbol{\beta}^\top \mathbf{x}(t)$. In the force-encoding model, $\mathbf{x}(t)$ was a two-element vector containing force and its first derivative. In the motor-unit-encoding model, $\mathbf{x}(t)$ contained the firing rates of all simultaneously recorded motor units. Encoder models were fit to single-trial data, which meant that only a subset of all neurons (173) could be fit to motor units. For the force-encoding model, 95% of all neurons had an $R^2 < 0.2$ (**fig. 3.3a**). For neurons recorded simultaneously with motor units, R^2 values were significantly larger for the motor-unit-encoding model than the force model ($p = 0.007$, t-test), but were also mostly low (**fig. 3.3b**; 95% < 0.2). Thus, single M1 neurons did

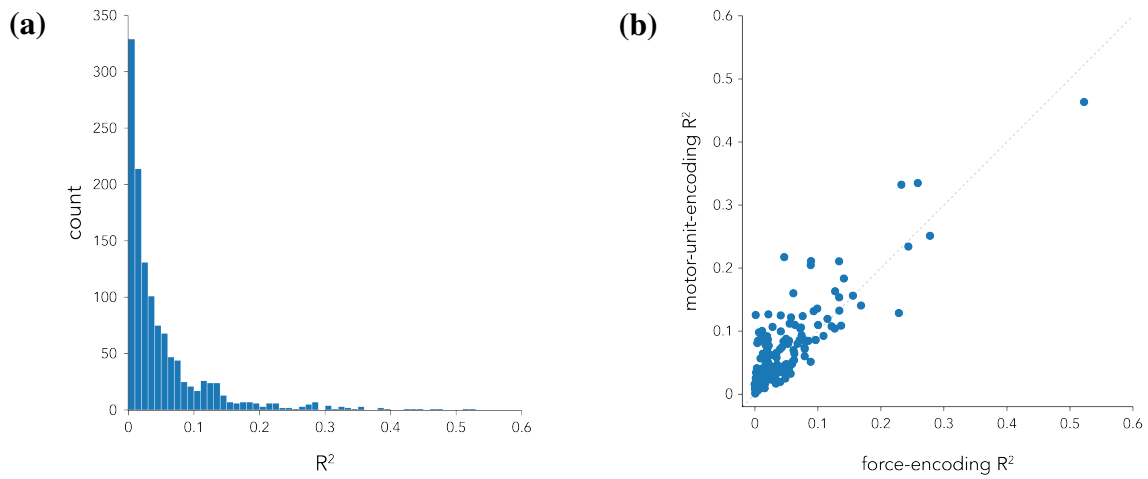


Figure 3.3: Encoder models of single-neuron responses. (**a**) Distribution across neurons of the R^2 when each response was fit to force and its derivative. Marker color indicates the size of the regularization parameter used in ridge regression. (**b**) Scatter plot of the R^2 when each neuron was fit to force and its derivative or to motor unit responses. Each point corresponds to one M1 neuron that was simultaneously recorded with motor units isolated from the lateral triceps.

not encode motor output, either at the level of force or motor unit activities.

Individual neurons correlating weakly with motor outputs does not preclude those outputs from being linearly read out by the neural population, but those readout dimensions may be small. During cycling, muscle readout dimensions capture a small fraction of the variance in M1 (~10% of the variance captured by its two leading principal components)[31]. It is conceivable that in a one-dimensional, isometric task, readout dimensions capture a larger share of neural variance. To investigate, we used ridge regression to decode force, EMG or motor unit signals from neural activity (pooling data across sessions). In principle, downstream circuits could decode behavior from neural activity using one of many readout dimensions, each carrying different strengths and weaknesses. A readout dimension might rely only on a small number of neurons whose activity patterns most closely resemble the behavior. Such readouts would optimally predict the behavior, but would be less robust to noise or injury in the few select neurons contributing to the readout. On the other hand, a readout dimension might try to leverage information from many neurons, potentially sacrificing predictive power for robustness. Mathematically, these options relate to the

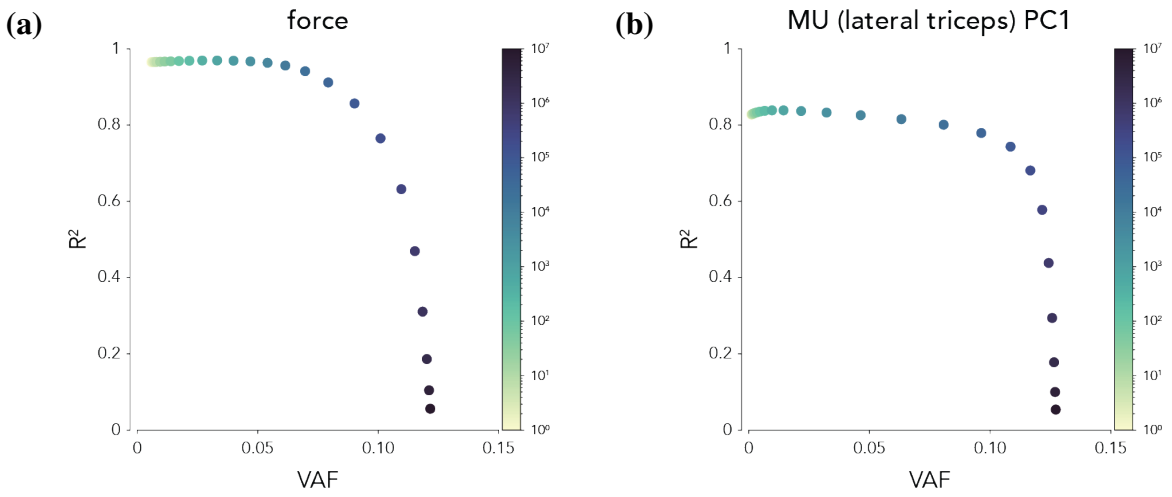


Figure 3.4: Bias-variance tradeoff in readout dimensions. (**a**) Generalization R^2 (10-fold cross validated) for predicting force amplitude from neural activity versus the proportion of neural variance explained by the force readout dimension. (**b**) Same as ***a***, but for predicting the first principal component in triceps MU activity.

bias-variance tradeoff in regularized linear regression[151]. We explored this trade off by computing the generalization performance of each decoder (10-fold cross validated) and the neural variance explained by the readout dimension over a range of regularization parameters. For small regularization parameters, mean force could be accurately decoded, but the decode dimension captured an infinitesimally small proportion of neural variance (**fig. 3.4a**; 0.58%). Increasing the regularization parameter increased the variance captured by the decode dimension, but caused the generalization performance to plummet. Moreover, the amount of variance captured only increased modestly (12.16%). Similar results were obtained when decoding the first principal component (PC) of MU activity in the triceps (**fig. 3.4b**), deltoid (**figs. S3.10b** and **S3.10c**), or multi-muscle EMG signals (**fig. S3.10a**). Thus, linear readout dimensions necessarily capture a small fraction of neural variance in M1, even during isometric force production.

3.2.4 Population structure

We explored the structure of population activity by projecting trial-averaged neural activity into its PC space. Projecting activity into the first three PCs revealed several notable features. The neural states during the three static force conditions were arranged roughly along a force-magnitude

axis (**fig. 3.5**, *bottom left, blue traces*). Specifically, the neural state during the intermediate static force condition was situated between the neural states during the low and high static force conditions. And the neural state during the low static force condition was closer to the initial state during other conditions that started at zero force. Despite this consistent arrangement with respect to the behavior, the neural states associated with all static force conditions occupied a small region of the space defined by the first three PCs, further reflecting that the largest signals in the population were not related to force. Also apparent in the leading PC space was that neural state during sinusoidal conditions (**fig. 3.5**, *bottom left, warm-colored traces*) separated from static activity along PC 1, but displayed little clear structure otherwise. Yet clearer rotational structure during sinusoidal conditions did emerge in PCs 5 and 6 (*bottom right*). Since rotations are typically one of, if not the largest signals in neural activity[31,155], the fact that rotational structure was buried in the 5th and 6th PCs was unexpected. This could have one of two interpretations: either rotations are not a dominant feature of neural activity during isometric force production or they occur in a subspace that is misaligned with neural activity from other conditions.

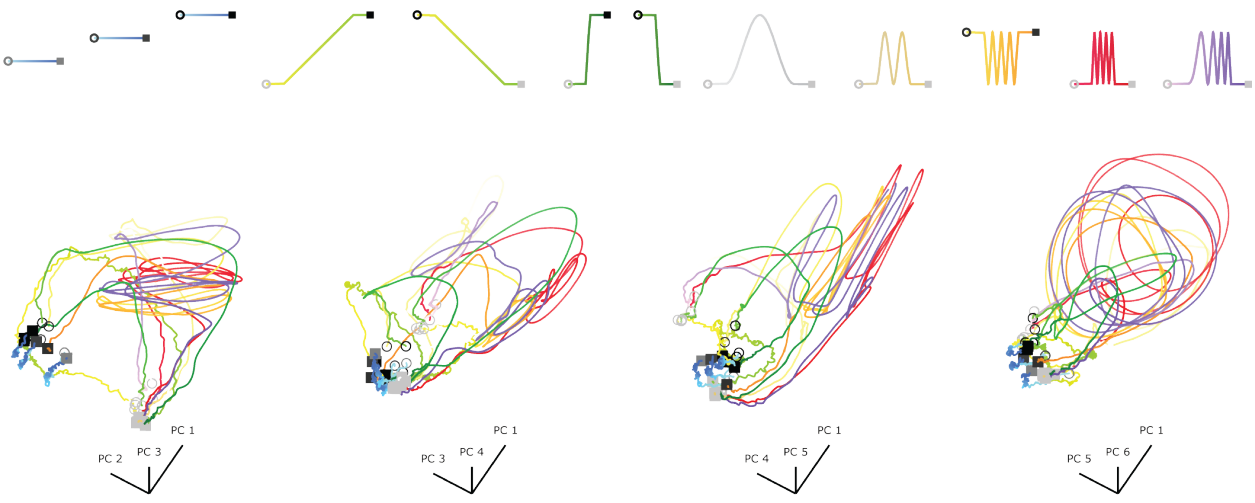


Figure 3.5: M1 activity visualized in a low-dimensional subspace. Principal components (PCs) were obtained from the trial-averaged firing rates of all neurons across all conditions. Condition force profiles are displayed on the *top row*. Condition start/end is indicated by *open circles/closed squares*, with endpoint force amplitude indicated by symbol brightness (e.g., *light open circles* indicate a low-force start). Neural activity was projected onto three of the six leading PCs (*bottom row*), with neural trajectories displayed using the same convention as the force profiles.

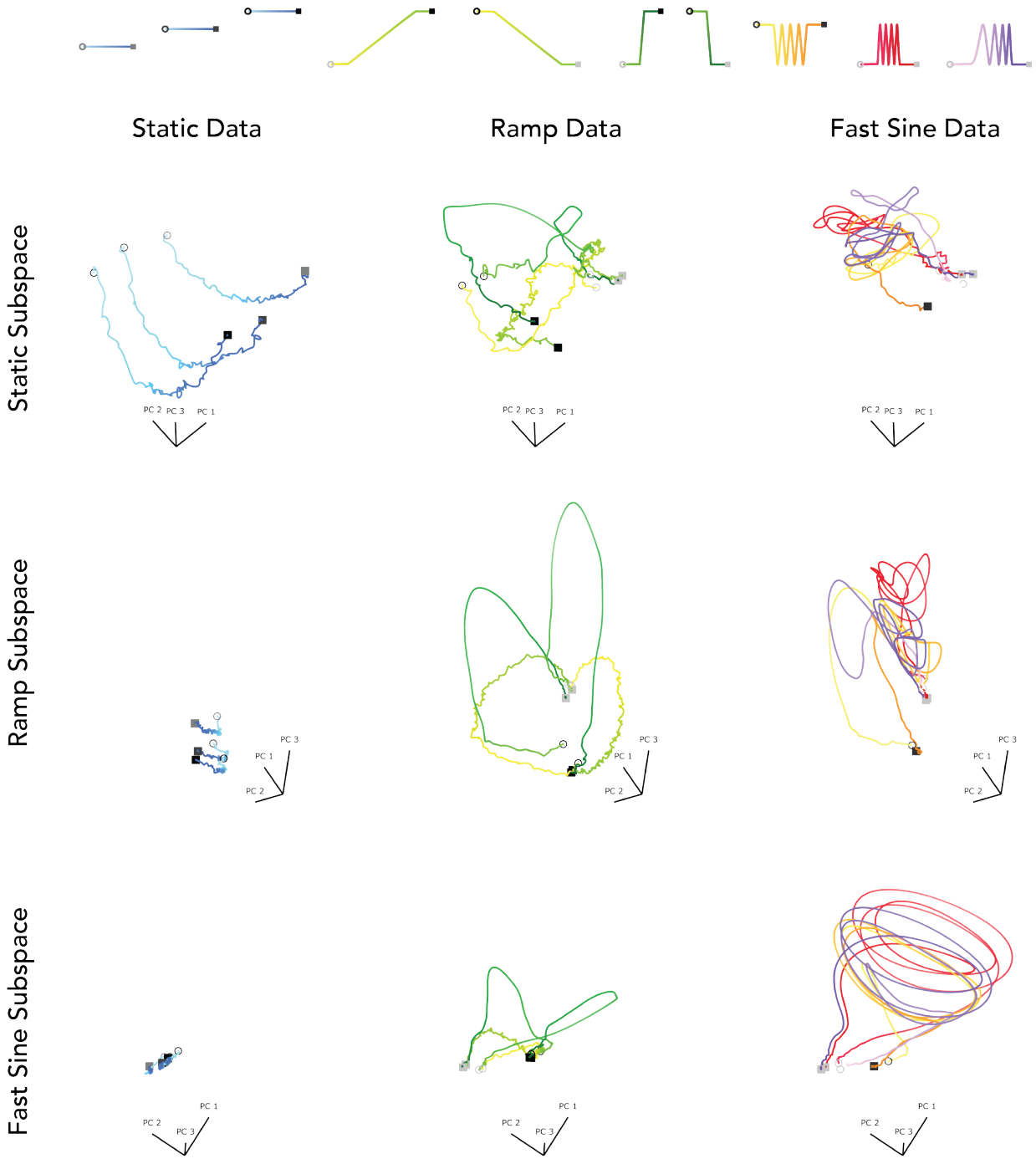


Figure 3.6: M1 activity in three different subspaces. The *top row* displays condition force profiles using the same convention as **fig. 3.5**. PCs were obtained from the trial-averaged firing rates of all neurons across one of three sets of conditions: 'static' (*blue traces*), 'ramp' (*green traces*), and 'fast sine' (*warm-colored traces*). Neural activity from each condition set (*columns*) was projected onto its first three PCs or those of the other condition sets (*rows*).

To further dissect population structure within and across conditions, we visualized neural activity in one of three subspaces defined by non-overlapping groups of conditions. These condition groups included the three static conditions, four ramp conditions, and three fastest sinusoidal conditions (2 Hz, 3 Hz and 0-3 Hz chirp). We then projected activity from each condition set into its own subspace or those defined by the other groups of conditions. When data from each set ('static', 'ramp', or 'fast sine') was projected into the top 3 PCs of its own subspace (**fig. 3.6**, *diagonal*), robust features emerged that were either diminished or entirely absent in the top 3 PCs of the full-condition subspace (**fig. 3.5**, *bottom left*). Static data were again arranged according to force amplitude (**fig. 3.6**, *top left*); data during slow and fast ramps rotated in the PC 1-2 plane and separated by speed along PC 3 (*center*); and fast sine data exhibited a condition-invariant translation and were ordered by speed along PC 1 and displayed robust rotations in the PC 2-3 plane (*bottom right*). In contrast, much of this structure was lost when either data set was projected into the top 3 PCs of the other two groups (*off diagonals*). Static data occupied a small portion of the ramp and fast sine subspaces, and the rotational structure and arrangement by speed or frequency were no longer observable in the ramp or fast sine data when projected into foreign subspaces.

3.2.5 Neural subspace alignment

The discrepancies between population-level features of neural activity projected into its own space or that associated with different conditions reveals a reorganization of neural coactivation patterns across conditions. **Fig. 3.6** indicates that such reorganization occurs within the largest (i.e., highest variance) dimensions. To quantify this reorganization, we measured the alignment between the neural subspaces for various groups of conditions, including but not limited to those shown in **fig. 3.6**. Given two data sets, A and B, the alignment index quantifies the amount of variance in A explained by the first 10 PCs of B, normalized by the total variance explained by the first 10 PCs of A[32]. **Fig. 3.7** shows the alignment indices across conditions groups; *yellow squares* indicate subspaces that are completely aligned, while *blue squares* indicate orthogonal subspaces. Note that some comparisons involved overlapping sets of conditions (e.g., sinusoidal versus fast

sinusoidal conditions; see Methods for a description of each condition set). The alignment indices revealed that the static subspace was nearly orthogonal to the subspace for every other set of conditions, most of all that including fast sinusoidal conditions. Yet other subspaces were also poorly aligned (alignment index < 0.5), such as the fast sinusoidal and slow ramp spaces or the slow sinusoidal and fast spaces. These results indicate that the dominant patterns of neural activity vary across conditions.

The alignment index succinctly quantifies the similarity between the largest neural signals, but does not consider the remaining smaller signals. It could be that neural activity quickly converges towards greater similarity beyond the 10th PC; or the small dimensions could be just as misaligned across conditions as the largest dimensions. To dissociate between these possibilities, we computed the cumulative proportion of total variance in one set of data captured by its PCs or those of other sets. 90% of the variance in static (**fig. 3.8**, *left*), ramp (*center*), or fast sine (*right*) data were

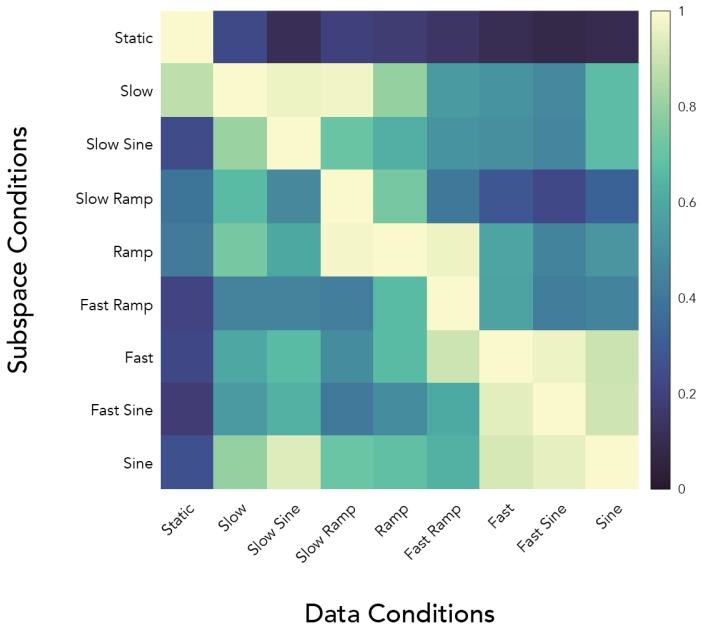


Figure 3.7: Subspace alignment across conditions. Alignment index (Elsayed *et al.*[32]) computed for trial-averaged neural activity across different sets of conditions. Each square indicates the amount of variance in 'data' (column) explained by the first PCs of 'subspace' (row), normalized by the amount of variance explained by the first 10 PCs of 'data'.

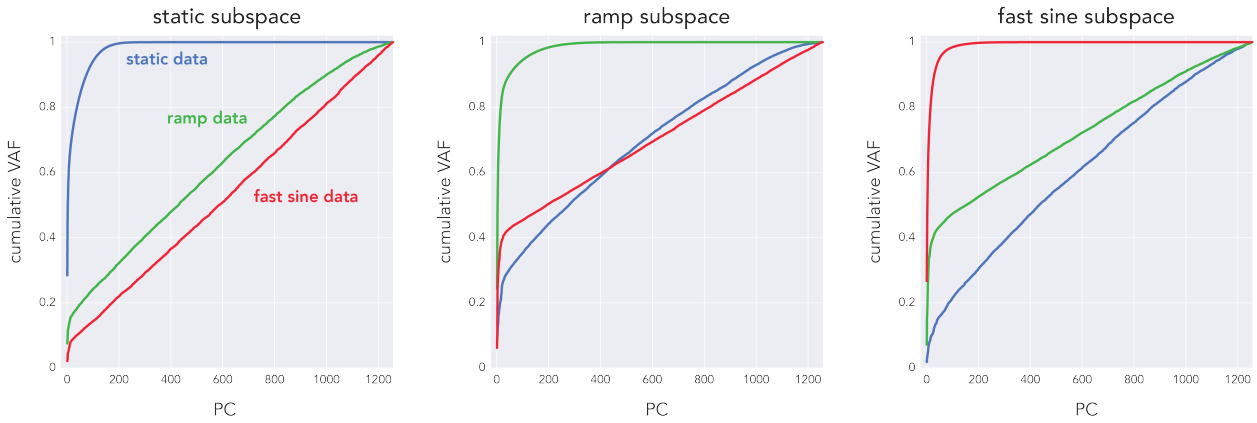


Figure 3.8: Variance explained across conditions. Cumulative proportion of variance in static (*blue*), ramp (*green*), or fast sine (*red*) data explained by each PC obtained from static (*left*), ramp (*center*), or fast sine (*right*) data.

captured by their leading 74, 48, and 29 PCs, respectively. In contrast, the cumulative variance in the ramp or fast sine data captured by the static subspace grew approximately linearly with each PC (**fig. 3.8**, *left*), indicating that the structure of neural covariance completely realigned between the static and other condition groups. Similar results were observed when quantifying the variance in static data captured by the fast sine subspace (**fig. 3.8**, *right*). Beyond the first ~20 PCs, the variance in static or fast sine data captured by the ramp subspace also grew close to linearly (**fig. 3.8**, *center*). In contrast, MU activity was much better aligned across conditions (**fig. S3.11**). Taken together, these findings indicate that most of the structure in neural activity undergoes at least three substantial reorganizations across conditions within this task.

3.3 Discussion

We found that generating isometric forces with various temporal profiles engages multiple, unrelated neural modes in M1. Prominent features in the population activity during one set of conditions (static, linear ramp, or fast sinusoidal force profiles) were not observed when activity was projected into the subspace obtained from a different set of conditions. On the farthest ends of the spectrum, neural activity unfolded in orthogonal subspaces during static and fast sinusoidal force production. These findings help contextualize our previous report that M1 contains hundreds

of neural dimensions in this task (**chapter 2**), which likely reflects the extensive reorganization of neural activity across different behaviors.

Despite the ubiquitous ability of tens or fewer signals to capture most of the neural variance during movement[26,32,35,36] and other cortical processes[6,37], several lines of evidence suggest that motor cortex may not necessarily be confined to a low-dimensional manifold. Motor cortical activity unfolds in orthogonal subspaces during the planning and execution of ballistic reaches[26,32] and during sustained rhythmic movements of one arm versus the other[158]. Additionally, during posture or movement tasks, the response gain of individual M1 neurons changes randomly with externally applied loads[159]. This was interpreted as indicating that the brain uses specialized processes for posture and motor control, which accords with our finding that neural activity occupies orthogonal subspaces during the generation of static or rapidly oscillating muscle forces. Beyond the motor system, recent work challenges the perception that cortex is confined to a low-dimensional manifold. In recordings of 10,000 neurons in mouse visual cortex, neural activity occupies a multi-dimensional space whose variance scales as a power law with the number of visual input stimuli[124]. These findings reveal that a large number of neural modes underlying cortical processes can emerge under the appropriate circumstances.

What are the necessary conditions for high-dimensional neural activity in motor cortex? And, more fundamentally, why does such structure emerge in the first place? The power-law scaling observed in the visual system reflects the proportion of neural variance allocated to representing coarse and fine stimulus features. It was argued that this balance is tuned to maximize the efficiency (by increasing dimensions dedicated to fine features) and smoothness (by increasing dimensions dedicated to coarse features) of the neural code[124]. This interpretation accords with recent findings that illuminate the functional importance of the largest (highest variance) dimensions in motor cortex. Neural smoothness can be measured with trajectory tangling; high tangling occurs when a particular neural state leads to dissimilar states at future time points[31]. Tangling can be reduced

by adding large signals that provide a scaffolding for smaller dimensions. In simulations, neural networks trained to have low tangling possess greater noise robustness relative to more highly tangled networks. And empirically, tangling in motor cortex is characteristically low during reaching and cycling[30,31]. We also observed that tangling was consistently much lower in M1 relative to motor units from one muscle (**fig. S3.12a**) or EMG signals from multiple muscles (**fig. S3.12b**). The consistently low tangling across a range of behaviors – isometric muscle contractions, ballistic reaches, and rhythmic cycling – indicates that preserving neural smoothness for the sake of noise robustness may be a significant driver for increasing the size or number of large dimensions to generate different behaviors. This may explain why the largest dimensions were nearly or entirely orthogonal across dissimilar behaviors, but does not address why the behaviors required so many different patterns of neural activity in the first place.

The neural manifold associated with a particular task can be conceptualized as an embedding of the task parameter manifold in neural space[37]. Consequently, neural dimensionality is fundamentally limited by that of the task. Task dimensionality can be mathematically described as depending on the autocorrelation length of each parameter, including but not limited to the duration of each behavioral condition[38]. Intuitively, neural activity can only explore as many dimensions as time permits. Time almost certainly underlies some of the complexity observed in the present task. Relative to point-to-point reaches, which unfold over several hundred ms[24,35], our conditions varied from 2-6 seconds in duration. Yet the cycling task also employed long conditions[31], so time cannot be the only factor at work. Given that the greatest disparity in neural activity was observed between static and high-frequency sinusoidal conditions, it is tempting to suggest that movement speed or frequency directly underlies cortical complexity. Neural response patterns also vary with speed during reaching[22], but individual neurons are not fit particularly well to movement kinematics[22], nor were they well fit to muscle force or its derivative (**fig. 3.3a**). Motor cortex may instead switch between different feedback control policies, which govern the form of its internal dynamics. The particular control policy could conceivably depend on any combination of the form

of task and feedback inputs, goal-related constraints imposed by the task, and some measure of behavioral complexity. Resolving these matters is paramount to fully understanding the breadth and depth of cortical control of movement. Given that we observed hundreds of neural modes during one-dimensional isometric force control, we may have only scratched the surface of the neural manifolds that facilitate diverse and flexible behaviors.

3.4 Methods

3.4.1 Data acquisition

Subject and task

All protocols were in accord with the National Institutes of Health guidelines and approved by the Columbia University Institutional Animal Care and Use Committee. Subject C was an adult, male macaque monkey (*Macaca mulatta*) weighing 13 kg.

The monkey performed the Pac-Man Task, as described in detail previously (**chapter 2**). Briefly, the monkey sat in a primate chair with his head restrained via surgical implant and his right arm loosely restrained. To perform the task, he grasped a manipulandum connected to a load cell on a ball bearing carriage mounted on a guide rail. Pac-Man was displayed on an LCD monitor at a fixed horizontal position. The monkey exerted force against the manipulandum to control Pac-Man's vertical position and intercept a scrolling dot path, which cued the temporal force profile that the monkey needed to produce to receive a juice reward.

We trained the monkey to generate static, step, ramp, and sinusoidal forces over a range of amplitudes and frequencies. We define a 'condition' as a particular target force profile (e.g., a 2 Hz sinusoid) that was presented on many 'trials', each a repetition of the same profile. Each condition included a 'lead-in' and 'lead-out' period: a one-second static profile appended to the beginning and end of the target profile, which facilitated trial alignment and averaging (see below). Trials

lasted 2.25-6 seconds, depending on the particular force profile. Juice was given throughout the trial so long as Pac-Man successfully intercepted the dots, with a large 'bonus' reward given at the end of the trial.

Target forces ranged from 0-16 Newtons. We employed twelve conditions presented interleaved in pseudo-random order: a random order was chosen, all conditions were performed, then a new random order was chosen. Three conditions employed static target forces: 33%, 66% and 100% of maximal force. Four conditions employed ramps: 0-to-16 or 16-to-0 Newtons, either fast (lasting 250 ms) or slow (lasting 4 s). Four conditions involved sinusoids at 0.25, 1, 2, and 3 Hz. The final condition was a 0-3 Hz chirp. All sinusoidal and chirp forces ranged from 0-12 Newtons, except for the 0.25 Hz sinusoid, which ranged from 0-16 Newtons.

Surgical procedures

After task performance stabilized at a high level, we performed a sterile surgery to implant a cylindrical chamber (Crist Instrument Co., 19 mm inner diameter) that provided access to M1. Guided by structural magnetic resonance imaging scans taken prior to surgery, we positioned the chamber surface-normal to the skull, centered over the central sulcus. We covered the skull within the cylinder with a thin layer of dental acrylic. Small (3.5 mm), hand-drilled burr holes through the acrylic provided the entry point for electrodes.

Cortical recordings

Neural activity was recorded in each session with S-Probes (Plexon) or Neuropixels probes. S-Probes contained 32 electrode sites with 100 m separation between them. Neuropixels probes contained 128 channels (two columns of 64 sites). Probes were lowered into position with a motorized microdrive (Narishige). Recordings were made at depths ranging from 5.6 - 13 mm relative to the surface of the dura. Raw neural signals were digitized at 30 kHz and saved with a 128-channel neural signal processor (Blackrock Microsystems, Cerebus).

EMG recordings

In multiple sessions, intramuscular EMG activity was recorded acutely from one muscle using closely spaced, modified paired hook-wire electrodes (Natus Neurology, PN 019-475400). Electrodes were modified to create quadrifilar electrodes by threading two pairs of wires into one needle, as described previously (**chapter 2**). Four quadrifilar electrodes were inserted ~ 1 cm into the muscle belly using 30 mm x 27 G needles. Needles were promptly removed and only the wires remained in the muscle during recording. Wires were thin (50 μm diameter) and flexible and their presence in the muscle is typically not felt after insertion, allowing the task to be performed normally. Wires were removed at the end of the session. Recordings were made from the anterior and lateral heads of the deltoid and the lateral head of the triceps.

In an additional session, intramuscular EMG was recorded acutely from eight muscles of the upper arm: superior trapezius, sternal pectoralis, deltoid (anterior and lateral heads), triceps (lateral and long heads), and biceps (short and long heads). Recordings were made using standard paired hook-wire electrodes (Natus Neurology, PN 019-475400) inserted percutaneously into each muscle. Wires were removed at the end of the session.

Raw voltages were amplified and analog filtered (band-pass 10 Hz - 10 kHz) with ISO-DAM 8A modules (World Precision Instruments), then digitized at 30 kHz with a neural signal processor (Blackrock Microsystems, Cerebus). EMG signals were digitally band-pass filtered online (50 Hz - 5 kHz) and saved.

3.4.2 Data processing

Signal processing and spike sorting

Cortical voltage signals were spike sorted using KiloSort 2.0[130]. A total of 1257 neurons were isolated across 31 sessions. For the single-muscle recording sessions, EMG signals were digitally

filtered offline using a second-order 500 Hz high-pass Butterworth. Any low SNR or dead EMG channels were omitted from analyses. Motor unit (MU) spike times were extracted using a custom semi-automated algorithm, as described previously (**chapter 2**). For the multi-muscle recording session, EMG signals were rectified and smoothed with a 25 ms Gaussian kernel.

Trial alignment and averaging

Single-trial spike rasters, for a given neuron or MU, were converted into a firing rate via convolution with a 25 ms Gaussian kernel. To facilitate trial averaging, trials for a given condition were aligned temporally and the average firing rate, at each time, was computed across trials. Stimulation trials were simply aligned to stimulation onset. For all other conditions, each trial was aligned on the moment the target force profile 'began' (when the target force profile reached Pac-Man). This alignment brought the actual (generated) force profile closely into register across trials. However, because the actual force profile could sometimes slightly lead or lag the target force profile, some modest across-trial variability remained. Thus, for all trials with changing forces, we realigned each trial to minimize the mean squared error between the actual force and the target force profile. This ensured that trials were well-aligned in terms of the actual generated forces (the most relevant quantity for analyses of MU activity). Trials were excluded from analysis if they could not be well aligned despite searching over shifts from -200 to 200 ms.

3.4.3 Data Analysis

Encoder models

For the force-encoding model, the firing rate of M1 neuron i was modeled as

$$r_i^{\mathrm{M1}}(t-\tau) = \beta_1 f(t) + \beta_2 \dot{f}(t) \tag{3.1}$$

where $f(t)$ is the force and $\dot{f}(t)$ is its first derivative.

For the motor-unit-encoding model, the firing rate of M1 neuron i was modeled as

$$r_i^{\mathrm{M1}}(t-\tau) = \sum_{j=1}^{m} \beta_j r_j^{\mathrm{MU}}(t) \tag{3.2}$$

where r_j^{MU} is the firing rate of the j^{th} MU that was simultaneously recorded with neuron i.

Parameter weights for both models were fit to single-trial data using ordinary least squares. Prior to fitting, covariates and responses were mean-centered across trials. τ was optimized over $[-200, 200]$ ms in 20 ms steps. **Figs. 3.3a** and **3.3b** show the fit R^2 for the optimal τ,

Preprocessing

Prior to the following population-level analyses, data were mean centered across conditions. M1 and MU activities were then soft normalized with a softening factor of 5, and smoothed EMG signals fully normalized, as has been used previously[29,31,158].

Bias-variance tradeoff in readout dimensions

For all data variables (M1 or MU responses, EMG signals, and forces), single-trial data was arranged as $X^{s,c} \in \mathbb{R}^{T_c \times N_s \times L_{s,c}}$, where T_c is the number of samples in condition c, N_s is the number of "units" (neurons, MUs, EMG channels, or 1 for the forces) recorded in session s, and $L_{s,c}$ is the number of trials for condition c in session s. To facilitate assessing model generalization using data pooled across sessions, each $X^{s,c}$ was used to construct $X_k^{s,c} \in \mathbb{R}^{T_c \times N_s}$ ($k = 1, 2 \ldots, \kappa$) by averaging over every κ^{th} trial, starting with trial k (κ was set to 10 to create 10 folds). For each k, $X_k^{s,c}$ was stacked vertically (across conditions) and horizontally (across units) to yield $X_k \in \mathbb{R}^{(\sum_c T_c) \times N}$, where $N = \sum_s N_s$.

Ridge regression was used to find a readout dimension

$$\boldsymbol{\beta}_k = (R_{!k}^\top R_{!k} + \lambda \mathbf{1})^{-1} R_{!k}^\top \mathbf{y}_{!k} \tag{3.3}$$

where $R_{!k} \in \mathbb{R}^{((\kappa-1)*\sum_c T_c)\times N}$ is the vertical concatenation of all but the k^{th} trial-averaged fold of M1 responses (i.e., X_k for M1 responses). The response $\mathbf{y}_{!k} \in \mathbb{R}^{((\kappa-1)*\sum_c T_c)\times 1}$ was either the session-averaged forces (i.e., averaging X_k across N, then vertically stacking all but the k^{th} fold) (**fig. 3.4a**), or PC 1 of lateral triceps MUs (i.e., projecting X_k onto its first PC, then vertically stacking all but the k^{th} fold projection) (**fig. 3.4b**), PC 1 of EMG channels (**fig. S3.10a**), PC 1 of anterior deltoid MUs (**fig. S3.10b**), or PC 1 of lateral deltoid (**fig. S3.10c**), mean centered across times. λ was logarithmically sampled between 1 and 10^7 in steps of 25. The predicted response for each fold was then computed as $\hat{\mathbf{y}}_k = R_k \boldsymbol{\beta}_k$.

The generalization R^2 was computed as

$$1 - \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2}{\|\mathbf{y}\|_2^2} \tag{3.4}$$

where $\mathbf{y} = [\mathbf{y}_1^\top, \mathbf{y}_2^\top, \ldots, \mathbf{y}_\kappa^\top]$ and $\hat{\mathbf{y}} = [\hat{\mathbf{y}}_1^\top, \hat{\mathbf{y}}_2^\top, \ldots, \hat{\mathbf{y}}_\kappa^\top]$.

The variance explained by the readout was computed as

$$\frac{\boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta}}{\mathrm{Tr}(\Sigma)\|\boldsymbol{\beta}\|_2^2} \tag{3.5}$$

where $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \ldots, \boldsymbol{\beta}_\kappa^\top]^\top \in \mathbb{R}^{\kappa N \times 1}$ and $\Sigma = \mathrm{cov}([R_1, R_2, \ldots, R_\kappa]^\top) \in \mathbb{R}^{\kappa N \times \kappa N}$.

Subspace alignment

The subspace alignment index (**fig. 3.7**) was computed as described in (Elsayed *et al.*)[32]:

$$A = \frac{\mathrm{Tr}\left(D_1^\top \Sigma_2 D_1\right)}{\sum_{k=1}^{10} \sigma_2(k)} \tag{3.6}$$

where D_1 is a $N \times 10$ matrix containing the first 10 principal components of the trial-averaged responses from the 'subspace condition set' (**fig. 3.7**, *rows*), Σ_2 is the $N \times N$ covariance matrix for the trial-averaged responses from 'data condition set' (**fig. 3.7**, *columns*), and $\sigma_2(k)$ is the k^{th} eigenvalue of Σ_2. The conditions included in each set were as follows:

- **static**: low, middle, and high static conditions
- **slow**: all statics, increasing and decreasing slow (4 s) ramps, and 0.25 Hz sinusoid
- **slow sine**: 0.25 and 1 Hz sinusoids
- **slow ramp**: increasing and decreasing slow ramps
- **ramp**: increasing and decreasing slow and fast (250 ms) ramps
- **fast ramp**: increasing and decreasing fast ramps
- **fast**: increasing and decreasing fast ramps; 1, 2, and 3 Hz sinusoids; and 0-3 Hz chirp
- **fast sine**: 2 and 3 Hz sinusoids and 0-3 Hz chirp
- **sine**: 0.25, 1, 2, and 3 Hz sinusoids; and 0-3 Hz chirp

Orthogonality across conditions

To assess the orthogonality of neural activity across conditions, PCA was applied to trial-averaged firing rates from the 'static', 'ramp', or 'fast sine' conditions (as defined above). The responses from each condition set were then projected onto the PCs from the same or different set and the variance in each projected dimension normalized by the total variance.

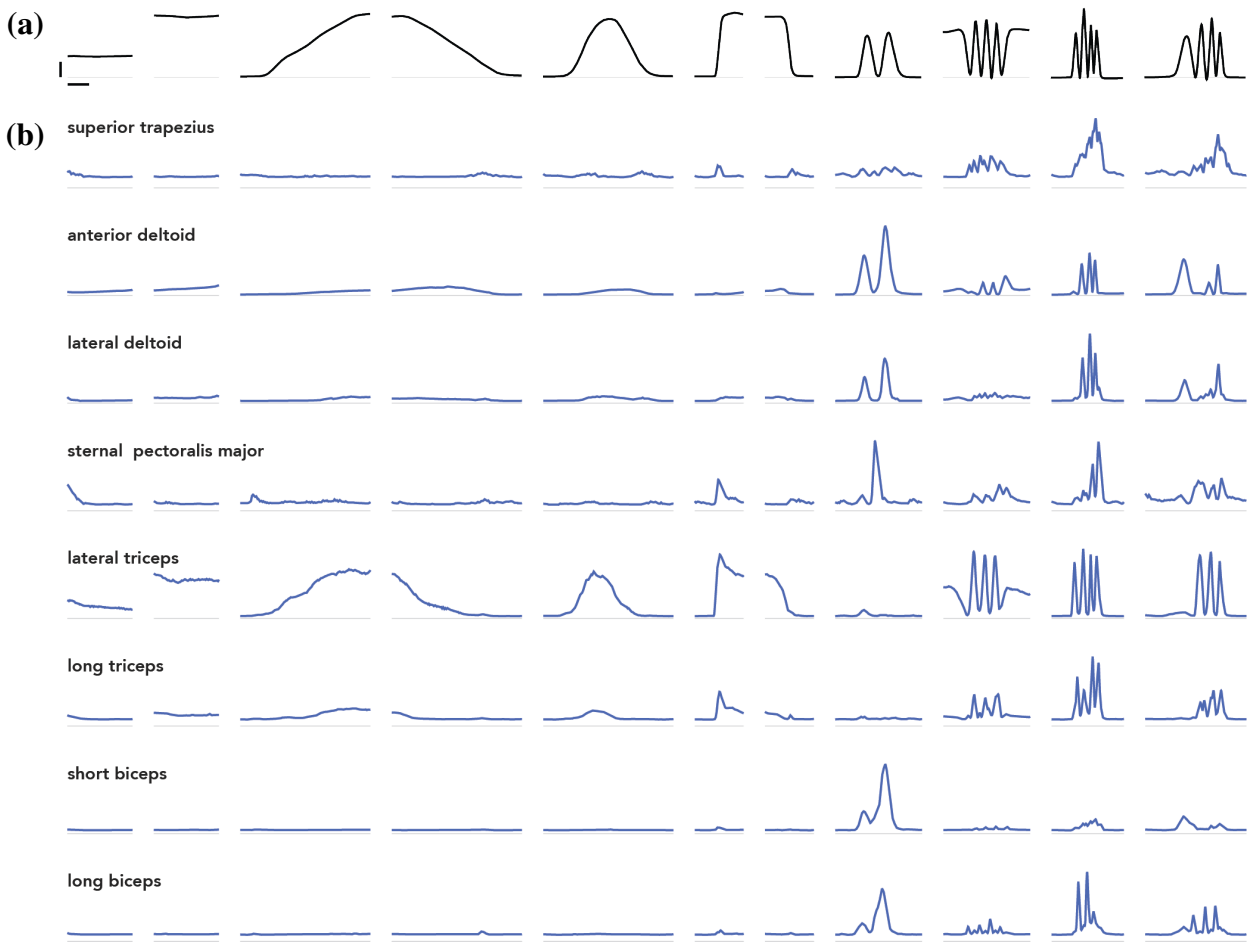
Tangling

Tangling (**fig. S3.12**) was computed as described in (Russo *et al.*)[31]:

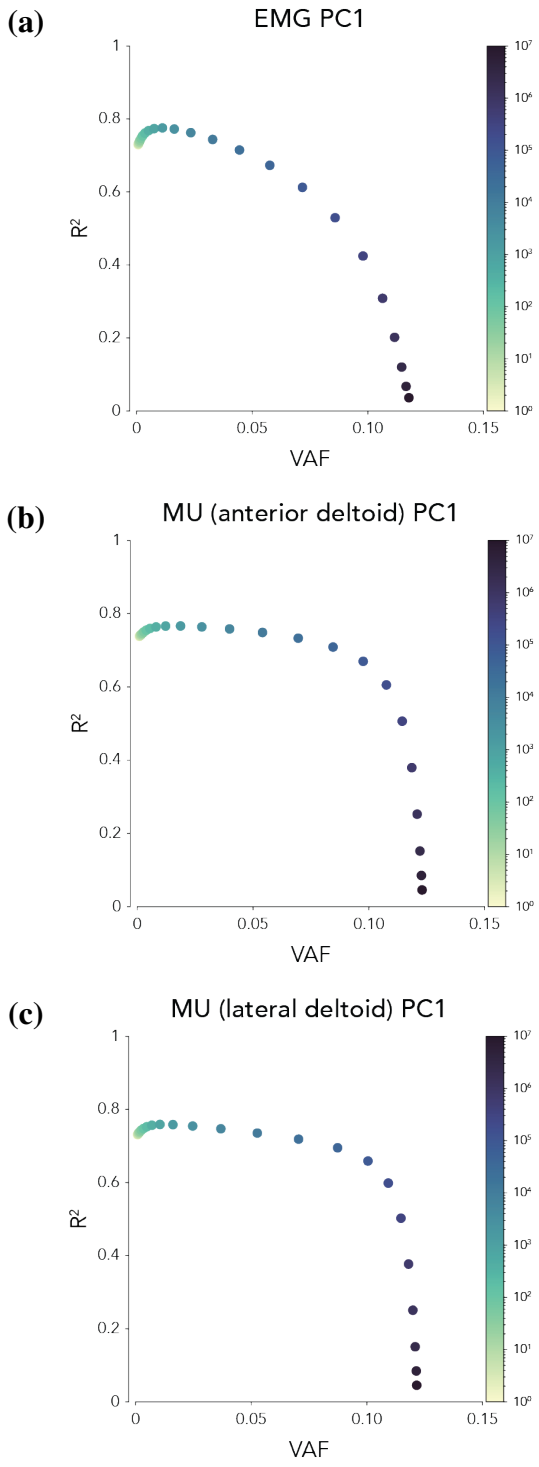
$$Q(t) = \max_{t'} \frac{\|\dot{\mathbf{x}}_t - \dot{\mathbf{x}}_{t'}\|^2}{\|\mathbf{x}_t - \mathbf{x}_{t'}\|^2 + \varepsilon} \tag{3.7}$$

where $\mathbf{x}_t$ was the 8-dimensional neural state at time t and $\dot{\mathbf{x}}_t$ is the derivative of the neural state. For computational efficiency, $\mathbf{x}$ was first downsampled to 250 Hz (every 4$^{\text{th}}$ sample). The constant $\varepsilon = 0.1 \cdot T^{-1} \sum_{t=1}^{T} \|\mathbf{x}_t\|^2$. For M1 and MUs, $\mathbf{x}$ was constructed by projected the responses onto their first 8 PCs.

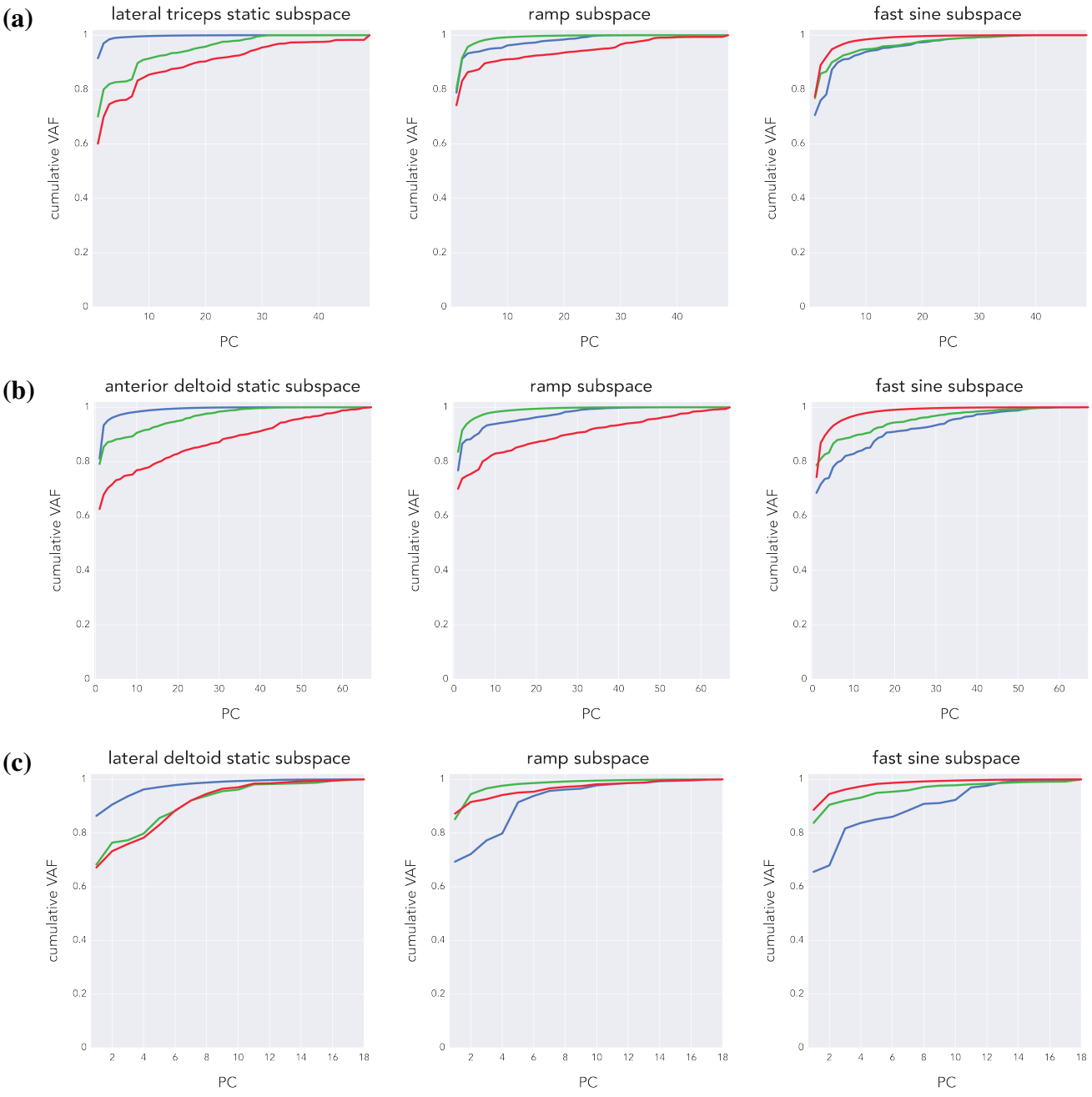
3.5 Supplementary Figures



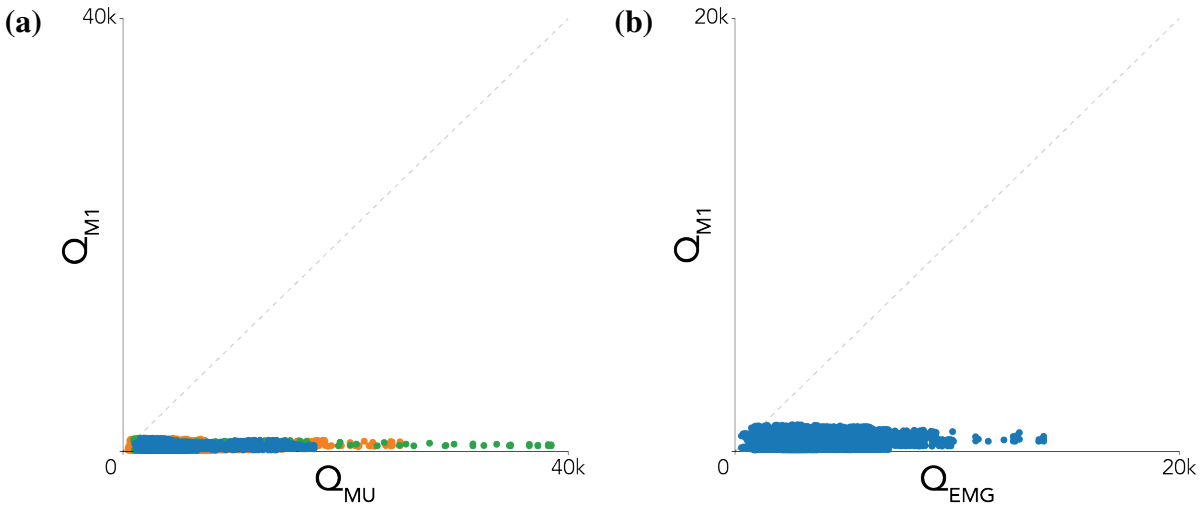
Supplementary Figure S3.9: Multi-muscle EMG activity. (**a**). Trial-averaged forces for 11 of 12 conditions (intermediate static force condition is omitted for space). Vertical scale bar indicate 8 N. Horizontal scale bar indicates 1 s. (**b**) Rectified and smoothed EMG signals recorded from 8 muscles of the upper arm. The EMG signals from each muscle were normalized across conditions.



Supplementary Figure S3.10: Bias-variance tradeoff in motor readouts. (**a**). Generalization R^2 (10-fold cross validated) for predicting the first principal component of multi-mucle EMG signals (**fig. S3.9b**) from neural activity versus the proportion of neural variance explained by the EMG component readout dimension. (**b**) Same as ***a***, but for the first principal component in anteriod deltoid MU activity. (**c**) Same as ***b***, but for the lateral deltoid.



Supplementary Figure S3.11: Motor unit variance explained across conditions. Cumulative proportion of variance in static (*blue*), ramp (*green*), or fast sine (*red*) data explained by each PC obtained from static (*left*), ramp (*center*), or fast sine (*right*) data. Data corresponds to trial-averaged firing rates of **(a)** lateral triceps, **(b)** anterior deltoid, or **(c)** lateral deltoid MUs.



Supplementary Figure S3.12: Trajectory tangling. (**a**) Trajectory tangling of M1 versus MUs from the lateral triceps (*green*), lateral deltoid (*orange*), and anterior deltoid (*blue*). (**b**) Trajectory tangling of M1 versus multi-channel EMG signals (**fig. S3.9b**).

Chapter 4: Conclusion

4.1 Summary

In this dissertation, I presented my work on the neural degrees of freedom underlying activity in the primary motor cortex (M1) and motor units (MUs) that control one muscle. In **chapter 2**, I presented an isometric force-tracking task (the 'Pac-Man Task') that enabled us to cue a monkey to generate muscle forces with diverse temporal profiles (static, ramp, and sinusoidal with constant or changing frequency). According to Henneman's size principle, MU activities should modulate in lockstep (i.e., varying along one degree of freedom), regardless of the context[68] (reviewed in **section 1.6**). MU responses adhered to this canonical description when forces were generated slowly with the arm in one posture, consistent with standard experimental conditions that supported the size principle[51,62,72–78,160]. Yet MU activity exhibited greater flexibility when forces changed swiftly or were generated with the arm in different postures. MUs also responded flexibly to supraspinal input. Intracortical microstimulation could substantially alter the patterns of MU coactivation and disrupt the recruitment order observed during steady force production. High-density recordings within M1 revealed hundreds of neural activity modes, at least an order of magnitude higher than previous estimates[26,32,35,36,123], demonstrating that cortical degrees of freedom were not a limiting factor in flexible MU control. We used a latent factor model to rigorously test how well the canonical description of MU control fit our data. When the factor model was limited to one degree of freedom, but free to infer any one-dimensional latent drive and learn any set of flexible MU link functions, it was unable to account for the diversity of MU responses, indicating a fundamental shortcoming of the hypothesis of rigid MU control (i.e., the size principle[45,68] and common drive[71]). Optimal MU recruitment strategies were predicted using a simple model of isometric force production by an idealized motor neuron pool that incorporated some

known diversity in MU size and speed (reviewed in **section 1.5**). Optimal recruitment strategies, based on the premise of minimizing total force error, agreed with the size principle for slow force profiles, but predicted more flexible MU activity patterns when forces changed swiftly, as observed empirically. These observations indicate that a broader optimality principle may guide MU control.

In **chapter 3**, I investigated neural activity in M1 and its relationship with behavior in the Pac-Man Task. In some respects, this task closely resembled those employed by Evarts in the first studies of M1, with the motor output varying along one dimension (i.e., force amplitude)[8,9] via isometric muscle contractions[9]. As Evarts reported that most M1 neurons resembled force amplitude or its derivative[8,9], it may not have been surprising to find similar results in the Pac-Man Task. Nevertheless, the firing rate of most neurons correlated poorly with force and its derivative across all conditions in the Pac-Man Task. M1 neurons correlated slightly stronger, albeit similarly weak overall, with the firing rate of simultaneously recorded MUs. Motor outputs – force or the first principal component of multi-muscle EMG signals or intramuscular MU activities – could be linearly read out from the M1 population, but readout dimensions captured very little variance in neural activity ($\leq 12\%$), even when aggressively regularized. These observations further indicated that M1 is poorly described by a 'representational model' (reviewed in **section 1.2**). Conversely, the population response exhibited rotational structure and low tangling – characteristics of the 'dynamical systems model' of M1[24,26–31]. But the central question of **chapter 3** was whether the hundreds of neural activity modes observed in M1 (reported in **chapter 2**) reflected meaningful organization in how it generates different behaviors, or simply reflected a better method to quantify neural modes. Low-dimensional projections of neural activity into the principal component space identified from responses during static, ramp, or fast sinusoidal conditions exhibited clear and robust structure, but that structure was lost when neural activity from one set of conditions was projected into the subspace associated with a different set of conditions. The largest signals during static and fast sinusoidal conditions were nearly orthogonal, as measured with an alignment index[32], and each pair of condition-specific subspaces were poorly aligned, as measured by the

cross-condition variance explained with principal component analysis. All together, these observations indicate that M1 leverages unrelated neural activity modes to generate different behaviors, in contrast with recently proposed low-dimensional theories of M1[6].

4.2 Future Directions

4.2.1 Optimal motor unit recruitment

The optimality hypothesis for MU control proposed in **chapter 2** requires additional evidence before definitive conclusions can be drawn. One of the most compelling lines of evidence would be to relate the contractile properties of empirically recorded MUs with their activity patterns in different contexts. For example, the optimal recruitment model (and general intuition) predicted that fast-twitch MUs are preferentially recruited when forces change quickly. But does this actually occur? Addressing this question requires some estimate of each MU's twitch response. One approach that has been employed during voluntary contractions involves taking a MU-spike-triggered average of the measured force[51,69]. However, the spike-triggered average (STA) distorts the estimated force and speed parameters of estimated twitch responses, even at relatively low discharge rates (5 spikes/s)[161]. Moreover, these distortions affect different types of MUs to an unequal degree, meaning that even conclusions based on relative contractile properties would likely be dubious. This consideration largely precluded investigating this question with my current data set, since most MUs that were preferentially active during rapid force conditions were not observed during static conditions. Another consideration is potential sources of contamination. In previous approaches, forces were generated by a single muscle (first dorsal interosseous)[51,69]. Yet it is clear that the Pac-Man Task engages multiple muscles (**fig. S3.9**), which may also distort STAs for MUs from one muscle. Alternative approaches to the STA include intraneural or intramuscular stimulation[162], which may necessitate investigating this question in man.

There are several ways in which the optimal recruitment model (**section 2.6.2**) could be extended. The current force model only incorporates one (agonist) muscle, but of course most joints

are crossed by multiple muscles. It also assumes an inverse relationship between MU size and speed (i.e., small MUs are slow and large MUs are fast), which is a reasonable approximation for some muscles[49,51], but does not always hold[162]. An additional consideration is that the distribution of fiber types differs widely across muscles. For example, the soleus contains primarily of slow-twitch fibers while the gastrocnemii contain more heterogeneous fiber types[163]. In fact, in cats, the soleus is more active than the lateral gastrocnemius during the maintenance of static posture, but this relationship reverses during rapid paw shakes[148]. Thus, flexible MU recruitment might be more pronounced in certain muscles, which may be predicted by a model that incorporates multiple diverse muscles. With regard to the optimization procedure, optimal recruitment in the current model is predicated on minimizing the total mean-squared error between the generated and desired force profile. This assumption may not be entirely unreasonable here, since the current form of the Pac-Man Task requires the monkey to constantly track the cued force profile, and he is encouraged to remain as close as possible to the underlying dot path (reward was delivered stochastically and scaled inversely with the distance between Pac-Man's and the target dots' centers). Yet a more general formulation might consider only minimizing task-relevant variability, as in optimal feedback control[153]. If cortex does mediate certain aspects of flexible MU recruitment, it may only do so when necessary or relevant for the task at hand. These predictions could also be investigated empirically via task modifications (see below). On the other hand, perhaps certain aspects of flexible recruitment are wired into the spinal circuitry and only emerge during certain behaviors that activate pathways not normally engaged during steady force production or postural maintenance, as suggested by Friedman[57] (see **section 1.4**). Incorporating feedback and recurrence to the motor neuron pool model, along with a more realistic cost function, may help guide future investigations to resolve the extent to which descending inputs or spinal mechanisms guide flexible MU recruitment.

4.2.2 Sources of high dimensionality in cortex

A central question raised by the results of **chapter 3** is what exactly is driving neural activity to be high dimensional? Possible explanations include the complexity of the visual stimuli or the need to control MUs in flexible manners. Preliminary modeling work in the lab, using deep networks trained to perform the Pac-Man Task in a similar manner as the monkey, indicates that the visual input may play a large role in driving high dimensional neural activity. Yet similar high dimensionality has not been observed in motor cortex during other tasks that use complex visual stimuli, such as cycling[31]. This suggests that the precise manner in which the visual stimuli are used to perform continuous tracking may be the root cause. One simple control to address this would be to include "catch" trials, wherein the same dot path scrolls across the screen, but the monkey is actively discouraged from generating any forces. More broadly, perhaps motor cortex leverages dramatically different feedback control policies depending on the demands of the task or context. This may underlie the observed orthogonality between neural activity while maintaining force statically or modulating force quickly. Investigating these questions thoroughly may require additional task modifications, as discussed below.

4.2.3 Task elaborations

I believe that the Pac-Man Task is ripe for exploration. I will refrain from detailing specific hypotheses, but will simply note several aspects of the task that can be readily tuned and how they might relate to particular areas of interest. Considering fewer, but longer conditions could facilitate the study of fatigue. Some evidence indicates that the characteristic feature of fatigue – namely, a reduction in the maximal force capacity of a muscle – involves supraspinal and spinal mechanisms[44]. At least one obvious complication with studying the neural basis of fatigue is the inability to extract large trial counts to facilitate trial averaging, as commonly precedes neural analyses. Yet recent advancements in high-density electrodes capable of recording thousands of neurons within a session[130], combined with methods for extracting meaningful population-level signals on single trials[164], might mean that the optimal time to study the neural mechanisms of fatigue is rapidly

approaching.

Another potential area of interest relates to optimal feedback control. As described above, this framework predicts that an optimal controller only aims to correct task-relevant variability[153]. As also described, the current version of the task encourages uniform variability over the full course of the dot path. This constraint could be relaxed by, for example, replacing the circular target dots with vertically oriented ellipses and varying their curvature, either continuously within a condition or across conditions. Other readily tunable parameters include the force gain; Pac-Man and/or target dot jitter, size, and opacity; target scroll speed; and the presence of obstacles. These parameters might also be used to investigate feedback control and/or motor adaptation. Undoubtedly, there are several interesting directions one could imagine.

References

[1] O. Gredal, H. Pakkenberg, M. Karlsborg, and B. Pakkenberg, "Unchanged total number of neurons in motor cortex and neocortex in amyotrophic lateral sclerosis: A stereological study," *Journal of Neuroscience Methods*, vol. 95, no. 2, 2000.

[2] S. H. Scott, "Inconvenient Truths about neural processing in primary motor cortex," *Journal of Physiology*, vol. 586, no. 5, 2008.

[3] F. Buchthal and H. Schmalbruch, *Motor unit of mammalian muscle*, 1980.

[4] C. L. Gooch, T. J. Doherty, K. M. Chan, M. B. Bromberg, R. A. Lewis, D. W. Stashuk, M. J. Berger, M. T. Andary, and J. R. Daube, *Motor unit number estimation: A technology and literature review*, 2014.

[5] M. T. Turvey, "Coordination.," *American Psychologist*, 1990.

[6] J. A. Gallego, M. G. Perich, L. E. Miller, and S. A. Solla, "Neural Manifolds for the Control of Movement," *Neuron*, vol. 94, no. 5, 2017.

[7] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. J. Hudspeth, *Principles of Neural Science*, 5th ed. McGraw-Hill Medical, 2013.

[8] E. V. Evarts, "Relation of pyramidal tract activity to force exerted during voluntary movement.," *Journal of neurophysiology*, vol. 31, no. 1, 1968.

[9] ——, "Activity of pyramidal tract neurons during postural fixation.," *Journal of neurophysiology*, vol. 32, no. 3, 1969.

[10] P. D. Cheney and E. E. Fetz, "Functional classes of primate corticomotoneuronal cells and their relation to active force," *Journal of Neurophysiology*, vol. 44, no. 4, 1980.

[11] E. V. Evarts, C. Fromm, J. Kroller, and V. A. Jennings, "Motor cortex control of finely graded forces," *Journal of Neurophysiology*, vol. 49, no. 5, 1983.

[12] W. T. Thach, "Correlation of neural discharge with pattern and force of muscular activity, joint position, and direction of intended next movement in motor cortex and cerebellum," *Journal of Neurophysiology*, vol. 41, no. 3, 1978.

[13] C. Fromm, "Changes of steady state activity in motor cortex consistent with the length-tension relation of muscle," *Pflügers Archiv European Journal of Physiology*, vol. 398, no. 4, 1983.

[14] A. Riehle and J. Requin, "Monkey primary motor and premotor cortex: Single-cell activity related to prior information about direction and extent of an intended movement," *Journal of Neurophysiology*, vol. 61, no. 3, 1989.

[15] A. P. Georgopoulos, J. F. Kalaska, R. Caminiti, and J. T. Massey, "On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex," *Journal of Neuroscience*, vol. 2, no. 11, 1982.

[16] A. P. Georgopoulos, A. B. Schwartz, and R. E. Kettner, "Neuronal population coding of movement direction," *Science*, vol. 233, no. 4771, 1986.

[17] J. F. Kalaska, D. A. Cohen, M. L. Hyde, and M. Prud'homme, "A comparison of movement direction-related versus load direction-related activity in primate motor cortex, using a two-dimensional reaching task," *Journal of Neuroscience*, vol. 9, no. 6, 1989.

[18] J. Ashe and A. P. Georgopoulos, "Movement parameters and neural activity in motor cortex and area 5," *Cerebral Cortex*, vol. 4, no. 6, 1994.

[19] Q. G. Fu, D. Flament, J. D. Coltz, and T. J. Ebner, "Temporal encoding of movement kinematics in the discharge of primate primary motor and premotor neurons," *Journal of Neurophysiology*, vol. 73, no. 2, 1995.

[20] D. W. Moran and A. B. Schwartz, "Motor cortical representation of speed and direction during reaching," *Journal of Neurophysiology*, vol. 82, no. 5, 1999.

[21] R. Caminiti, P. B. Johnson, C. Galli, S. Ferraina, and Y. Burnod, "Making arm movements within different parts of space: The premotor and motor cortical representation of a coordinate system for reaching to visual targets," *Journal of Neuroscience*, vol. 11, no. 5, 1991.

[22] M. M. Churchland and K. V. Shenoy, "Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex," *Journal of Neurophysiology*, vol. 97, no. 6, 2007.

[23] E. Todorov, "Direct cortical control of muscle activation in voluntary arm movements: A model," *Nature Neuroscience*, vol. 3, no. 4, 2000.

[24] M. M. Churchland, J. P. Cunningham, M. T. Kaufman, J. D. Foster, P. Nuyujukian, S. I. Ryu, K. V. Shenoy, and K. V. Shenoy, "Neural population dynamics during reaching," *Nature*, vol. 487, no. 7405, 2012.

[25] K. V. Shenoy, M. Sahani, and M. M. Churchland, "Cortical control of arm movements: A dynamical systems perspective," *Annual Review of Neuroscience*, vol. 36, 2013.

[26] M. T. Kaufman, M. M. Churchland, S. I. Ryu, and K. V. Shenoy, "Cortical activity in the null space: Permitting preparation without movement," *Nature Neuroscience*, vol. 17, no. 3, 2014.

[27] C. Pandarinath, V. Gilja, C. H. Blabe, P. Nuyujukian, A. A. Sarma, B. L. Sorice, E. N. Eskandar, L. R. Hochberg, J. M. Henderson, and K. V. Shenoy, "Neural population dynamics in human motor cortex during movements in people with ALS," *eLife*, vol. 4, no. JUNE, 2015.

[28] J. C. Kao, P. Nuyujukian, S. I. Ryu, M. M. Churchland, J. P. Cunningham, and K. V. Shenoy, "Single-trial dynamics of motor cortex and their applications to brain-machine interfaces," *Nature Communications*, vol. 6, 2015.

[29] A. H. Lara, J. P. Cunningham, and M. M. Churchland, "Different population dynamics in the supplementary motor area and motor cortex during reaching," *Nature Communications*, vol. 9, no. 1, 2018.

[30] A. K. Suresh, J. M. Goodman, E. V. Okorokova, M. T. Kaufman, N. G. Hatsopoulos, and S. J. Bensmaia, "Neural population dynamics in motor cortex are different for reach and grasp," *eLife*, vol. 9, 2020.

[31] A. A. Russo, S. R. Bittner, S. M. Perkins, J. S. Seely, B. M. London, A. H. Lara, A. Miri, N. J. Marshall, A. Kohn, T. M. Jessell, L. F. Abbott, J. P. Cunningham, and M. M. Churchland, "Motor Cortex Embeds Muscle-like Commands in an Untangled Population Response," *Neuron*, vol. 97, no. 4, 2018.

[32] G. F. Elsayed, A. H. Lara, M. T. Kaufman, M. M. Churchland, and J. P. Cunningham, "Reorganization between preparatory and movement population responses in motor cortex," *Nature Communications*, vol. 7, no. 1, 2016.

[33] D. Sussillo, M. M. Churchland, M. T. Kaufman, and K. V. Shenoy, "A neural network that finds a naturalistic solution for the production of muscle activity," *Nature Neuroscience*, vol. 18, no. 7, 2015.

[34] J. A. Michaels, B. Dann, and H. Scherberger, "Neural Population Dynamics during Reaching Are Better Explained by a Dynamical System than Representational Tuning," *PLoS Computational Biology*, vol. 12, no. 11, 2016.

[35] J. A. Gallego, M. G. Perich, S. N. Naufel, C. Ethier, S. A. Solla, and L. E. Miller, "Cortical population activity within a preserved neural manifold underlies multiple motor behaviors," *Nature Communications*, vol. 9, no. 1, 2018.

[36] P. T. Sadtler, K. M. Quick, M. D. Golub, S. M. Chase, S. I. Ryu, E. C. Tyler-Kabara, B. M. Yu, and A. P. Batista, "Neural constraints on learning," *Nature*, vol. 512, no. 7515, 2014.

[37] P. Gao and S. Ganguli, *On simplicity and complexity in the brave new world of large-scale neuroscience*, 2015.

[38] P. Gao, E. Trautmann, B. M. Yu, G. Santhanam, S. Ryu, K. Shenoy, and S. Ganguli, "A theory of multineuronal dimensionality, dynamics and measurement," *bioRxiv*, 2017.

[39] B. M. Yu, C. Kemere, G. Santhanam, A. Afshar, S. I. Ryu, T. H. Meng, M. Sahani, and K. V. Shenoy, "Mixture of trajectory models for neural decoding of goal-directed movements," *Journal of Neurophysiology*, vol. 97, no. 5, 2007.

[40] S. J. Wood and C. R. Slater, *Safety factor at the neuromuscular junction*, 2001.

[41] L. Landmesser, "The distribution of motoneurones supplying chick hind limb muscles.," *The Journal of Physiology*, vol. 284, no. 1, 1978.

[42] R. N. Lemon, *Descending pathways in motor control*, 2008.

[43] J. A. Rathelot and P. L. Strick, "Subdivisions of primary motor cortex based on cortico-motoneuronal cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 3, 2009.

[44] U. Windhorst, *Muscle proprioceptive feedback and spinal networks*, 2007.

[45] E. Henneman, G. Somjen, and D. O. Carpenter, "Functional significance of cell size in spinal motoneurons," *Journal of neurophysiology*, vol. 28, 1965.

[46] R. M. Enoka and A. J. Fuglevand, *Motor unit physiology: Some unresolved issues*, 2001.

[47] S. C. Bodine, R. R. Roy, E. Eldred, and V. R. Edgerton, "Maximal force as a function of anatomical features of motor units in the cat tibialis anterior," *Journal of Neurophysiology*, vol. 57, no. 6, 1987.

[48] R. B. Stein, A. S. French, A. Mannard, and R. Yemm, "New methods for analysing motor function in man and animals," *Brain Research*, vol. 40, no. 1, 1972.

[49] A. J. Fuglevand, D. A. Winter, and A. E. Patla, "Models of recruitment and rate coding organization in motor-unit pools," *Journal of Neurophysiology*, vol. 70, no. 6, 1993.

[50] F. Buchthal and H. Schmalbruch, "Contraction Times and Fibre Types in Intact Human Muscle," *Acta Physiologica Scandinavica*, vol. 79, no. 4, 1970.

[51] J. A. Stephens and T. P. Usherwood, "The mechanical properties of human motor units with special reference to their fatiguability and recruitment threshold," *Brain Research*, vol. 125, no. 1, 1977.

[52] D. Pette and R. S. Staron, "Transitions of muscle fiber phenotypic profiles," *Histochemistry and Cell Biology*, vol. 115, no. 5, 2001.

[53] D. Pette and G. Vrbová, "Invited review: Neural control of phenotypic expression in mammalian muscle fibers," *Muscle & Nerve*, vol. 8, no. 8, 1985.

[54] A. J. Buller, J. C. Eccles, and R. M. Eccles, "Interactions between motoneurones and muscles in respect of the characteristic speeds of their responses," *The Journal of Physiology*, vol. 150, no. 2, 1960.

[55] D. Pette and G. Vrbová, "What does chronic electrical stimulation teach us about muscle plasticity," *Muscle and Nerve*, vol. 22, no. 6, 1999.

[56] L. Malisoux, M. Francaux, and D. Theisen, "What do single-fiber studies tell us about exercise training?" *Medicine and Science in Sports and Exercise*, vol. 39, no. 7, 2007.

[57] W. A. Friedman, G. W. Sypert, J. B. Munson, and J. W. Fleshman, "Recurrent inhibition in type-identified motoneurons," *Journal of Neurophysiology*, vol. 46, no. 6, 1981.

[58] B. E. Tomlinson and D. Irving, "The numbers of limb motor neurons in the human lumbosacral cord throughout life," *Journal of the Neurological Sciences*, vol. 34, no. 2, 1977.

[59] M. H. Schieber, "Comparative anatomy and physiology of the corticospinal system," in *Handbook of Clinical Neurology*, vol. 82, 2007, ch. 2.

[60] D. W. Stashuk, "EMG signal decomposition: How can it be accomplished and used?" *Journal of Electromyography and Kinesiology*, vol. 11, no. 3, 2001.

[61] D. Denny-Brown and D. Phil, "On the nature of postural reflexes," *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character*, vol. 104, no. 730, 1929.

[62] E. D. Adrian and D. W. Bronk, "The discharge of impulses in motor nerve fibres: Part II. The frequency of discharge in reflex and voluntary contractions.," *The Journal of physiology*, vol. 67, no. 2, 1929.

[63] E. Henneman, "Relation between size of neurons and their susceptibility to discharge," *Science*, vol. 126, no. 3287, 1957.

[64] E. Henneman, G. Somjen, and D. O. Carpenter, "Excitability and inhibitability of motoneurons of different sizes," *Journal of neurophysiology*, vol. 28, no. 3, 1965.

[65] C. B. Olson, D. O. Carpenter, and E. Henneman, "Orderly Recruitment of Muscle Action Potentials: Motor Unit Threshold and EMG Amplitude," *Archives of Neurology*, vol. 19, no. 6, 1968.

[66] H. P. Clamann and E. Henneman, "Electrical measurement of axon diameter and its use in relating motoneuron size to critical firing level," *Journal of Neurophysiology*, vol. 39, no. 4, 1976.

[67] G. Somjen, D. O. Carpenter, and E. Henneman, "Responses of motoneurons of different sizes to graded stimulation of supraspinal centers of the brain.," *Journal of neurophysiology*, vol. 28, no. 5, 1965.

[68] E. Henneman, H. P. Clamann, J. D. Gillies, and R. D. Skinner, "Rank order of motoneurons within a pool: law of combination," *Journal of Neurophysiology*, vol. 37, no. 6, 1974.

[69] H. S. Milner-Brown, R. B. Stein, and R. Yemm, "The orderly recruitment of human motor units during voluntary isometric contractions," *The Journal of Physiology*, vol. 230, no. 2, 1973.

[70] C. J. De Luca, R. S. LeFever, M. P. McCue, and A. P. Xenakis, "Control scheme governing concurrently active human motor units during voluntary contractions," *The Journal of Physiology*, vol. 329, no. 1, 1982.

[71] C. J. De Luca and Z. Erim, *Common drive of motor units in regulation of muscle force*, 1994.

[72] R. S. Person and L. P. Kudina, "Discharge frequency and discharge pattern of human motor units during voluntary contraction of muscle," *Electroencephalography and Clinical Neurophysiology*, vol. 32, no. 5, 1972.

[73] A. W. Monster and H. Chan, "Isometric force production by motor units of extensor digitorum communis muscle in man," *Journal of Neurophysiology*, vol. 40, no. 6, 1977.

[74] C. J. De Luca, R. S. LeFever, M. P. McCue, and A. P. Xenakis, "Behaviour of human motor units in different muscles during linearly varying contractions," *The Journal of Physiology*, vol. 329, no. 1, 1982.

[75] S. Riek and P. Bawa, "Recruitment of motor units in human forearm extensors," *Journal of Neurophysiology*, vol. 68, no. 1, 1992.

[76] J. Y. Hogrel, "Use of surface EMG for studying motor unit recruitment during isometric linear force ramp," *Journal of Electromyography and Kinesiology*, vol. 13, no. 5, 2003.

[77] C. J. De Luca and E. C. Hostage, "Relationship between firing rate and recruitment threshold of motoneurons in voluntary isometric contractions," *Journal of Neurophysiology*, vol. 104, no. 2, 2010.

[78] Y. Lei, N. L. Suresh, W. Z. Rymer, and X. Hu, "Organization of the motor-unit pool for different directions of isometric contraction of the first dorsal interosseous muscle," *Muscle and Nerve*, vol. 57, no. 1, 2018.

[79] J. S. Carp and J. R. Wolpaw, "Motor Neurons and Spinal Control of Movement," in *Encyclopedia of Life Sciences*, 2010.

[80] V. F. Harrison and O. A. Mortensen, "Identification and voluntary control of single motor unit activity in the tibialis anterior muscle," *The Anatomical Record*, vol. 144, no. 2, 1962.

[81] J. V. Basmajian, "Control and training of individual motor units," *Science*, vol. 141, no. 3579, 1963.

[82] I. H. Wagman, D. S. Pierce, and R. E. Burger, "Proprioceptive influence in volitional control of individual motor units," *Nature*, vol. 207, no. 5000, 1965.

[83] M. Kato and J. Tanji, "Volitionally controlled single motor units in human finger muscles," *Brain Research*, vol. 40, no. 2, 1972.

[84] J. S. Thomas, E. M. Schmidt, and F. T. Hambrecht, "Facility of motor unit control during tasks defined directly in terms of unit behaviors," *Experimental Neurology*, vol. 59, no. 3, 1978.

[85] S. Illyés, "The Voluntary Control of Single Motor Unit Activity," *IFAC Proceedings Volumes*, 1977.

[86] E. Formento, P. Botros, and J. M. Carmena, "A non-invasive brain-machine interface via independent control of individual motor units," *bioRxiv*, 2021.

[87] A. Nardone, C. Romanò, and M. Schieppati, "Selective recruitment of high-threshold human motor units during voluntary isotonic lengthening of active muscles.," *The Journal of Physiology*, vol. 409, no. 1, 1989.

[88] E. F. Hodson-Tole and J. M. Wakeling, "Variations in motor unit recruitment patterns occur within and between muscles in the running rat (Rattus norvegicus)," *Journal of Experimental Biology*, vol. 210, no. 13, 2007.

[89] ——, "Motor unit recruitment patterns 1: Responses to changes in locomotor velocity and incline," *Journal of Experimental Biology*, vol. 211, no. 12, 2008.

[90] J. M. Wakeling, K. Uehli, and A. I. Rozitis, "Muscle fibre recruitment can respond to the mechanics of the muscle contraction," *Journal of the Royal Society Interface*, vol. 3, no. 9, 2006.

[91] B. M. ter Haar Romeny, J. J. Denier van der Gon, and C. C. Gielen, "Relation between location of a motor unit in the human biceps brachii and its critical firing levels for different tasks," *Experimental Neurology*, vol. 85, no. 3, 1984.

[92] U. Herrmann and M. Flanders, "Directional tuning of single motor units," *Journal of Neuroscience*, vol. 18, no. 20, 1998.

[93] V. Von Tscharner, "Intensity analysis in time-frequency space of surface myoelectric signals by wavelets of specified resolution," in *Journal of Electromyography and Kinesiology*, vol. 10, 2000.

[94] J. M. Wakeling and D. A. Syme, "Wave properties of action potentials from fast and slow motor units of rats," *Muscle and Nerve*, vol. 26, no. 5, 2002.

[95] V. Von Tscharner and B. Goepfert, "Estimation of the interplay between groups of fast and slow muscle fibers of the tibialis anterior and gastrocnemius muscle while running," *Journal of Electromyography and Kinesiology*, vol. 16, no. 2, 2006.

[96] D. Farina, *Counterpoint: Spectral properties of the surface EMG do not provide information about motor unit recruitment and muscle fiber type*, 2008.

[97] L. M. Mendell and E. Henneman, "Terminals of single Ia fibers: location, density, and distribution within a pool of 300 homonymous motoneurons.," *Journal of neurophysiology*, vol. 34, no. 1, 1971.

[98] G. E. Loeb, "Motoneurone task groups: Coping with kinematic heterogeneity," *Journal of Experimental Biology*, vol. VOL. 115, 1985.

[99] J. A. Hoffer, G. E. Loeb, N. Sugano, W. B. Marks, M. J. O'Donovan, and C. A. Pratt, "Cat hindlimb motoneurons during locomotion. III. Functional segregation in sartorius," *Journal of Neurophysiology*, vol. 57, no. 2, 1987.

[100] P. N. Bawa, K. E. Jones, and R. B. Stein, "Assessment of size ordered recruitment," *Frontiers in Human Neuroscience*, vol. 8, no. July, 2014.

[101] N. C. Holt, J. M. Wakeling, and A. A. Biewener, "The effect of fast and slow motor unit activation on whole-muscle mechanical performance: The size principle may not pose a mechanical paradox," *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, no. 1783, 2014.

[102] J. E. Desmedt and E. Godaux, "Fast motor units are not preferentially activated in rapid voluntary contractions in man," *Nature*, vol. 267, no. 5613, 1977.

[103] C. J. Heckman and R. M. Enoka, "Motor Unit," *Comprehensive Physiology*, 2012.

[104] R. Bottinelli and C. Reggiani, *Human skeletal muscle fibres: Molecular and functional diversity*, 2000.

[105] D. S. Gokhin, N. E. Kim, S. A. Lewis, H. R. Hoenecke, D. D. D'Lima, and V. M. Fowler, "Thin-filament length correlates with fiber type in human skeletal muscle," *American Journal of Physiology - Cell Physiology*, vol. 302, no. 3, 2012.

[106] L. Luo, *Principles of Neurobiology*. 2015.

[107] D. F. Feeney, F. G. Meyer, N. Noone, and R. M. Enoka, "A latent low-dimensional common input drives a pool of motor neurons: A probabilistic latent state-space model," *Journal of Neurophysiology*, vol. 118, no. 4, 2017.

[108] W. Senn, K. Wyler, H. P. Clamann, J. Kleinle, H. R. Lüscher, and L. Müller, "Size principle and information theory," *Biological Cybernetics*, vol. 76, no. 1, 1997.

[109] K. E. Jones, A. F. C. Hamilton, and D. M. Wolpert, "Sources of signal-dependent noise during isometric force production," *Journal of Neurophysiology*, vol. 88, no. 3, 2002.

[110] R. B. Stein, E. R. Gossen, and K. E. Jones, *Neuronal variability: Noise or part of the signal?* 2005.

[111] A. Holtermann, K. Roeleveld, P. J. Mork, C. Grönlund, J. S. Karlsson, L. L. Andersen, H. B. Olsen, M. K. Zebis, G. $\mathrm{Sj_1}gaard$, and K. $\mathrm{S_1}gaard$, "Selective activation of neuromuscular compartments within the human trapezius muscle," *Journal of Electromyography and Kinesiology*, vol. 19, no. 5, 2009.

[112] D. Borzelli, M. Gazzoni, A. Botter, L. Gastaldi, A. d'Avella, and T. M. Vieira, "Contraction level, but not force direction or wrist position, affects the spatial distribution of motor unit recruitment in the biceps brachii muscle," *European Journal of Applied Physiology*, vol. 120, no. 4, 2020.

[113] R. J. Wyman, I. Waldron, and G. M. Wachtel, "Lack of fixed order of recruitment in cat motoneuron pools," *Experimental Brain Research*, vol. 20, no. 2, 1974.

[114] U. Windhorst, T. M. Hamm, and D. G. Stuart, "On the function of muscle and reflex partitioning," *Behavioral and Brain Sciences*, vol. 12, no. 4, 1989.

[115] P. Christova, A. Kossev, and N. Radicheva, "Discharge rate of selected motor units in human biceps brachii at different muscle lengths," *Journal of Electromyography and Kinesiology*, vol. 8, no. 5, 1998.

[116] E. F. Hodson-Tole and J. M. Wakeling, *Motor unit recruitment for dynamic tasks: Current understanding and future directions*, 2009.

[117] R. Q. Quiroga, Z. Nadasdy, and Y. Ben-Shaul, "Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering," *Neural Computation*, vol. 16, no. 8, 2004.

[118] F. Franke, R. Q. Quiroga, A. Hierlemann, and K. Obermayer, "Bayes optimal template matching for spike sorting – combining fisher discriminant analysis with optimal filtering," *Journal of Computational Neuroscience*, vol. 38, no. 3, 2015.

[119] J. E. Chung, J. F. Magland, A. H. Barnett, V. M. Tolosa, A. C. Tooker, K. Y. Lee, K. G. Shah, S. H. Felix, L. M. Frank, and L. F. Greengard, "A Fully Automated Approach to Spike Sorting," *Neuron*, vol. 95, no. 6, 2017.

[120] M. D. Binder, F. R. Robinson, and R. K. Powers, "Distribution of effective synaptic currents in cat triceps surae motoneurons. VI. Contralateral pyramidal tract," *Journal of Neurophysiology*, vol. 80, no. 1, 1998.

[121] M. D. Binder, R. K. Powers, and C. J. Heckman, "Nonlinear Input-Output Functions of Motoneurons," *Physiology (Bethesda, Md.)*, vol. 35, no. 1, 2020.

[122] E. Archer, I. M. Park, L. Buesing, J. P. Cunningham, and L. Paninski, "Black box variational inference for state space models," *arXiv*, 2015.

[123] B. M. Yu, J. P. Cunningham, G. Santhanam, S. I. Ryu, K. V. Shenoy, and M. Sahani, "Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity," *Journal of Neurophysiology*, vol. 102, no. 1, 2009.

[124] C. Stringer, M. Pachitariu, N. A. Steinmetz, M. Carandini, and K. D. Harris, "High-dimensional geometry of population responses in visual cortex," *Nature*, vol. 571, no. 7765, 2019.

[125] S. A. Overduin, A. d'Avella, J. Roh, J. M. Carmena, and E. Bizzi, "Representation of muscle synergies in the primate brain," *Journal of Neuroscience*, vol. 35, no. 37, 2015.

[126] E. Henneman, "Voluntary control of human motor unit," *The motor system, Neurophysiology and muscle mechanism*, 1976.

[127] M. H. Schieber and G. Rivlis, "Partial reconstruction of muscle activity from a pruned network of diverse motor cortex neurons," *Journal of Neurophysiology*, vol. 97, no. 1, 2007.

[128] D. M. Griffin, D. S. Hoffman, and P. L. Strick, "Corticomotoneuronal cells are "functionally tuned"," *Science*, vol. 350, no. 6261, 2015.

[129] G. E. Loeb and C. Gans, *Electromyography for Experimentalists*. The University of Chicago Press, 1986.

[130] N. A. Steinmetz, C. Aydin, A. Lebedeva, M. Okun, M. Pachitariu, M. Bauza, M. Beau, J. Bhagat, C. Böhm, M. Broux, S. Chen, J. Colonell, R. J. Gardner, B. Karsh, D. Kostadinov, C. Mora-Lopez, J. Park, J. Putzeys, B. Sauerbrei, R. J. van Daal, A. Z. Vollan, M. Welkenhuysen, Z. Ye, J. Dudman, B. Dutta, A. W. Hantman, K. D. Harris, A. K. Lee, E. I. Moser, J. O'Keefe, A. Renart, K. Svoboda, M. Häusser, S. Haesler, M. Carandini, and T. D. Harris, "Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings," *bioRxiv*, 2020.

[131] C. J. De Luca, "Electromyography," in *Encyclopedia of Medical Devices and Instrumentation*, Hoboken, NJ, USA: John Wiley & Sons, Inc., Apr. 2006.

[132] F. Negro, S. Muceli, A. M. Castronovo, A. Holobar, and D. Farina, "Multi-channel intramuscular and surface EMG decomposition by convolutive blind source separation," *Journal of Neural Engineering*, vol. 13, no. 2, 2016.

[133] D. Carlson and L. Carin, *Continuing progress of spike sorting in the era of big data*, 2019.

[134] M. Pachitariu, N. A. Steinmetz, S. Kadir, M. Carandini, and K. D. Harris, "Fast and accurate spike sorting of high-channel count probes with KiloSort," in *Advances in Neural Information Processing Systems*, 2016.

[135] J. H. Lee, D. Carlson, H. Shokri, W. Yao, G. Goetz, E. Hagen, E. Batty, E. J. Chichilnisky, G. Einevoll, and L. Paninski, "YASS: Yet another spike sorter," in *Advances in Neural Information Processing Systems*, vol. 2017-Decem, 2017.

[136] M. S. Lewicki, "A review of methods for spike sorting: The detection and classification of neural action potentials," *Network: Computation in Neural Systems*, vol. 9, no. 4, 1998.

[137] G. D. Brown, S. Yamada, and T. J. Sejnowski, "Independent component analysis at the neural cocktail party," *Trends in Neurosciences*, vol. 24, no. 1, 2001.

[138] L. J. Goldberg and B. Derfler, "Relationship among recruitment order, spike amplitude, and twitch tension of single motor units in human masseter muscle," *Journal of Neurophysiology*, vol. 40, no. 4, 1977.

[139] J. F. Magland and A. H. Barnett, "Unimodal clustering using isotonic regression: ISO-SPLIT," Aug. 2015.

[140] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, 1999.

[141] J. A. Hartigan and P. M. Hartigan, "The Dip Test of Unimodality," *The Annals of Statistics*, vol. 13, no. 1, 1985.

[142] C. Pouzat, O. Mazor, and G. Laurent, "Using noise signature to optimize spike-sorting and to assess neuronal classification quality," *Journal of Neuroscience Methods*, vol. 122, no. 1, 2002.

[143] C. M. Harris and D. M. Wolpert, "Signal-dependent noise determines motor planning," *Nature*, vol. 394, no. 6695, 1998.

[144] E. Henneman and L. M. Mendell, "Functional Organization of Motoneuron Pool and its Inputs," in *Comprehensive Physiology*, 2011, pp. 423–507.

[145] I. E. Brown and G. E. Loeb, "Measured and modeled properties of mammalian skeletal muscle: IV. Dynamics of activation and deactivation," *Journal of Muscle Research and Cell Motility*, vol. 21, no. 1, 2000.

[146] A. J. Van Soest and L. J. Casius, "Which factors determine the optimal pedaling rate in sprint cycling," *Medicine and Science in Sports and Exercise*, vol. 32, no. 11, 2000.

[147] V. J. Caiozzo and K. M. Baldwin, "Determinants of work produced by skeletal muscle: Potential limitations of activation and relaxation," *American Journal of Physiology*, vol. 273, no. 3 PART 1, 1997.

[148] J. L. Smith, B. Betts, V. R. Edgerton, and R. F. Zernicke, "Rapid ankle extension during paw shakes: Selective recruitment of fast ankle extensors," *Journal of Neurophysiology*, vol. 43, no. 3, 1980.

[149] J. M. Wakeling and T. Horn, "Neuromechanics of muscle synergies during cycling," *Journal of Neurophysiology*, vol. 101, no. 2, 2009.

[150] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. 2001.

[151] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. 2009, vol. 2nd.

[152] R. R. Neptune and S. A. Kautz, "Muscle Activation and Deactivation Dynamics: The Governing Properties in Fast Cyclical Human Movement Performance?" *Exercise and Sport Sciences Reviews*, vol. 29, no. 2, 2001.

[153] E. Todorov and M. I. Jordan, "Optimal feedback control as a theory of motor coordination," *Nature Neuroscience*, vol. 5, no. 11, 2002.

[154] S. H. Scott, "Optimal feedback control and the neural basis of volitional motor control," *Nature Reviews Neuroscience*, vol. 5, no. 7, 2004.

[155] M. T. Kaufman, J. S. Seely, D. Sussillo, S. I. Ryu, K. V. Shenoy, and M. M. Churchland, "The largest response component in the motor cortex reflects movement timing but not movement type," *eNeuro*, vol. 3, no. 4, 2016.

[156] G. F. Elsayed and J. P. Cunningham, "Structure in neural population recordings: An expected byproduct of simpler phenomena?" *Nature Neuroscience*, vol. 20, no. 9, 2017.

[157] C. Pandarinath, D. J. O'Shea, J. Collins, R. Jozefowicz, S. D. Stavisky, J. C. Kao, E. M. Trautmann, M. T. Kaufman, S. I. Ryu, L. R. Hochberg, J. M. Henderson, K. V. Shenoy, L. F. Abbott, and D. Sussillo, "Inferring single-trial neural population dynamics using sequential auto-encoders," *Nature Methods*, vol. 15, no. 10, 2018.

[158] K. C. Ames and M. M. Churchland, "Motor cortex signals for each arm are mixed across hemispheres and neurons yet partitioned within the population response," *eLife*, vol. 8, 2019.

[159] I. Kurtzer, T. M. Herter, and S. H. Scott, "Random change in cortical load representation suggests distinct control of posture and movement," *Nature Neuroscience*, vol. 8, no. 4, 2005.

[160] H. S. Milner-Brown, R. B. Stein, and R. Yemm, "The contractile properties of human motor units during voluntary isometric contractions," *The Journal of Physiology*, vol. 228, no. 2, 1973.

[161] B. Calancie and P. Bawa, "Limitations of the spike-triggered averaging technique," *Muscle & Nerve*, vol. 9, no. 1, 1986.

[162] B. Bigland-Ritchie, A. J. Fuglevand, and C. K. Thomas, "Contractile properties of human motor units: Is man a cat?" *Neuroscientist*, vol. 4, no. 4, 1998.

[163] S. Schiaffino and C. Reggiani, "Fiber types in Mammalian skeletal muscles," *Physiological Reviews*, vol. 91, no. 4, 2011.

[164] J. P. Cunningham and B. M. Yu, "Dimensionality reduction for large-scale neural recordings," *Nature Neuroscience*, vol. 17, no. 11, 2014.